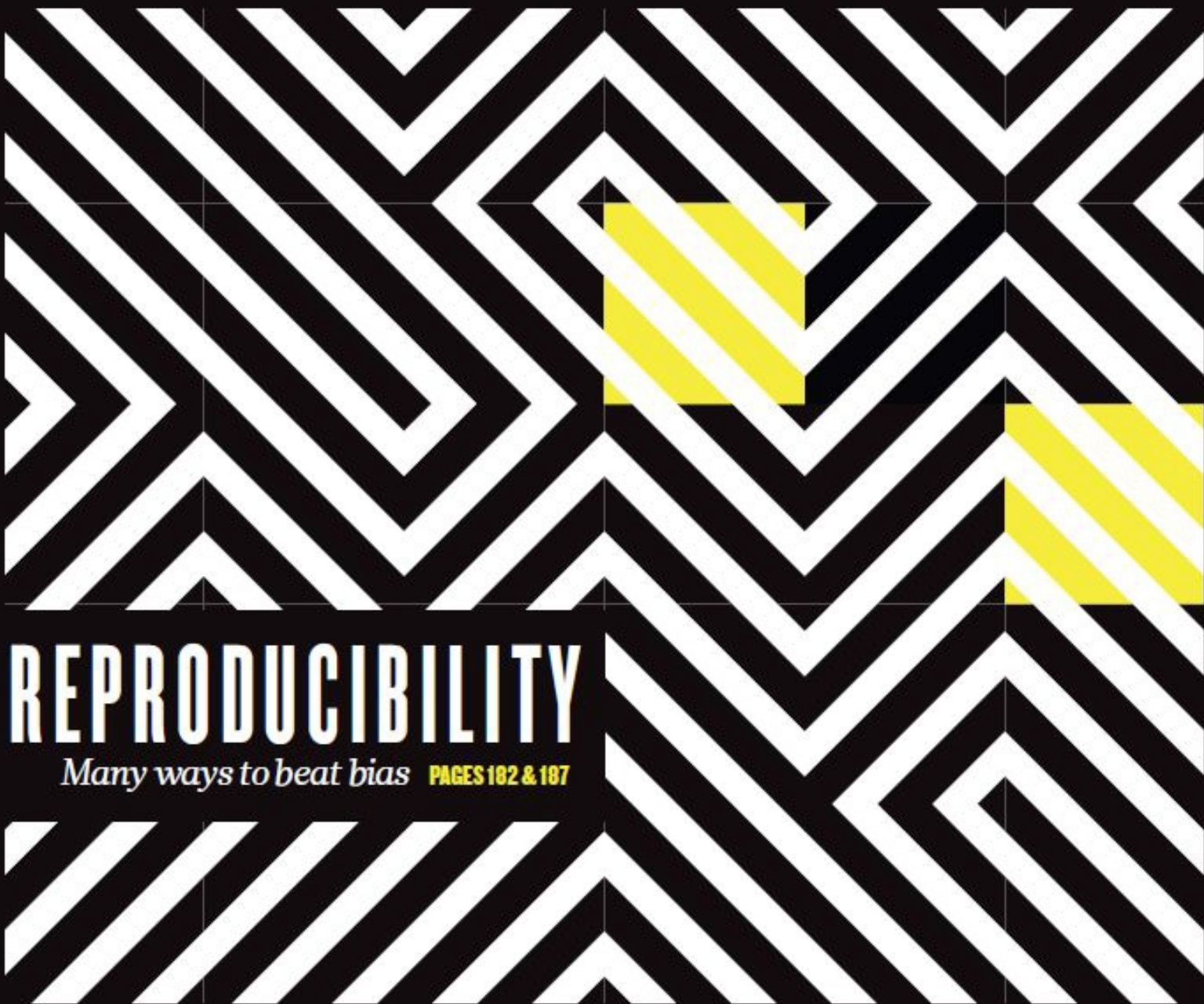


nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE



REPRODUCIBILITY

Many ways to beat bias **PAGES 182 & 187**

BIOGRAPHY

DOING THE CONTINENTAL

Alfred Wegener and the birth of plate tectonics

PAGE 182

EPIDEMIOLOGY

MALARIA: KEEP THE FAITH

The campaign continues in sub-Saharan Africa

PAGES 198 & 207

PLANETARY SCIENCE

CUTTING A DASH IN THE DUST

Fast-moving features in the AU Mic debris disk

PAGES 204 & 230

NATURE.COM/NATURE

8 October 2015 £10

Vol. 526, No. 7572

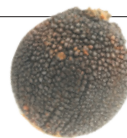


THIS WEEK

EDITORIALS

DRUGS New agreement to tackle pharmaceutical pollution **p.164**

WORLD VIEW Vaccination the best way to measure health care **p.165**



DUNG OVER Rolling beetles fooled by look-alike seeds **p.167**

Let's think about cognitive bias

The human brain's habit of finding what it wants to find is a key problem for research. Establishing robust methods to avoid such bias will make results more reproducible.

“Ever since I first learned about confirmation bias I’ve been seeing it everywhere.” So said British author and broadcaster Jon Ronson in *So You’ve Been Publicly Shamed* (Picador, 2015).

You will see a lot of cognitive bias in this week’s *Nature*. In a series of articles, we examine the impact that bias can have on research, and the best ways to identify and tackle it. One enemy of robust science is our humanity — our appetite for being right, and our tendency to find patterns in noise, to see supporting evidence for what we already believe is true, and to ignore the facts that do not fit.

The sources and types of such cognitive bias — and the fallacies they produce — are becoming more widely appreciated. Some of the problems are as old as science itself, and some are new: the IKEA effect, for example, describes a cognitive bias among consumers who place artificially high value on products that they have built themselves. Another common fallacy in research is the Texas sharp-shooter effect — firing off a few rounds and then drawing a bull’s eye around the bullet holes. And then there is asymmetrical attention: carefully debugging analyses and debunking data that counter a favoured hypothesis, while letting evidence in favour of the hypothesis slide by unexamined.

Such fallacies sound obvious and easy to avoid. It is easy to think that they only affect other people. In fact, they fall naturally into investigators’ blind spots (see page 182).

Advocates of robust science have repeatedly warned against cognitive habits that can lead to error. Although such awareness is essential, it is insufficient. The scientific community needs concrete guidance on how to manage its all-too-human biases and avoid the errors they cause.

That need is particularly acute in statistical data analysis, where some of the best-established methods were developed in a time before data sets were measured in terabytes, and where choices between techniques offer abundant opportunity for errors. Proteomics and genomics, for example, crunch millions of data points at once, over thousands of gene or protein variants. Early work was plagued by false positives, before the spread of techniques that could account for the myriad hypotheses that such a data-rich environment could generate.

Although problems persist, these fields serve as examples of communities learning to recognize and curb their mistakes. Another example is the venerable practice of double-blind studies. But more effort is needed, particularly in what some have called evidence-based data analysis: research on what techniques work best to establish default analytical pipelines for cleaning and debugging data sets, selecting models and other steps of analysis.

More specifically, science needs ways to identify the mistakes most likely to be made by novice (and not-so-novice) number crunchers. The scientific community must design research protocols that safeguard against these errors, and devise methods that ferret out sloppy analyses.

Some researchers already do this well, so one relatively simple strategy is to improve how knowledge and resources move from a narrow group of experts to the broader scientific community. If highly respected, easy-to-implement alternative routes are available and encouraged, it will be harder to cling to analyses that are rigged by conscious or unconscious bias to produce the results that researchers want. Funders should support teams that are attempting to determine the best analytical routes, and should provide training in data analysis for others. Institutions and principal investigators should make such training mandatory.

Finally, the scientific community must go beyond statistical safeguards, and improve researchers’ behaviour. Angst over unreliable research has already spurred investigations into ways to make results more robust. Some of the most promising address not just techniques, but also academic culture: laboratory and workplace habits can discourage rigour, or can enforce it through blinding, preregistering analytical plans, crowdsourcing analysis, formally laying out null and alternative hypotheses, and labelling analyses as exploratory or confirmatory.

Such strategies require effort, but offer significant rewards. Blind analysis forces creative thinking as researchers struggle to find explanations for hypothetical results. A Comment on page 187 explores these rewards and offers tips for researchers ready to try it.

“It is easy to think that fallacies only affect other people.”

Crowdsourcing shows how the same data set, analysed with different approaches, can yield a variety of answers; it is a reminder that single-team analysis is only part of the story. As a Comment on page 189 reveals, crowdsourced analyses and interdisciplinary projects can also compare analysis techniques across disciplines, and show how one field might hold lessons for another. Some differences in approach are probably down to cultural happenstance — “we have always done it this way” — rather than to selection of best practice. That should change.

To ensure that such practices actually strengthen science, scientists must subject the strategies themselves to scientific scrutiny. (No one should take recommendations to counter bias on faith!) Social scientists have an important role here — studies of science in action are essential. Careful observation of scientists can test which strategies are most effective under what circumstances, and can explore how debiasing strategies can best be integrated into routine scientific practice.

Funders should support efforts to establish the best methods of blind analysis, crowdsourcing and reviewing registered analysis plans, and should help meta-scientists to test and compare these practices. Ideally, the utility and burdens of these strategies under varying circumstances would be explored and published in the peer-reviewed literature. This information could then be fed into much-needed training programmes, and so better equip the next generation of scientists to do good science.

Finding the best ways to keep scientists from fooling themselves has so far been mainly an art form and an ideal. The time has come to make it a science. We need to see it everywhere. ■



NATURE.COM
For Nature’s special collection on reproducibility, see: go.nature.com/huhbyr

Time to get clean

Formal recognition of drug pollution will help to protect humans and ecosystems.

Most nations have strict controls on environmental waste, from arsenic to zinc. Yet no legal limits have been set to control pollution from drugs during their manufacture, use and disposal. That is despite evidence that pharmaceutical waste can wreak havoc in the environment — hormones found in contraceptives cause male fish to grow female sex organs, and a painkiller used in livestock has wiped out millions of vultures in India that fed on the carcasses.

The need for global action was recognized internationally for the first time last week at a meeting in Geneva, Switzerland, led by the United Nations Environment Programme. The move is a small but significant development.

Pharmaceuticals pollute the environment mainly because wastewater treatment plants do not adequately remove compounds found in the drugs that people ingest and excrete. High concentrations are also released into water during drug manufacture. Other pollution comes from unused medicines that have not been safely disposed of, particularly in developing countries where stockpiles of outdated donated medicines can build up and leach into the environment. The industry points to studies that find pharmaceutical pollution does not pose an immediate risk to human health, because the concentrations in drinking water are not high enough to cause problems. But the levels found in the environment still damage wildlife and ecosystems.

Last week, countries, the drug industry and non-governmental bodies formally agreed — for the first time — that humans and ecosystems need protection from pharmaceutical pollution. A resolution passed at the triennial International Conference on Chemicals Management

(ICCM) also backs the need for global cooperation to build awareness and push for action to address drug pollution. The deal puts the issue permanently on the ICCM's radar, and is a crucial first step towards building much-needed initiatives to address the problem.

The ICCM is a middleweight organization with high-level backing, and so it is able to make an impact. A large part of its remit is to keep an eye on progress towards a voluntary goal to ensure that, by 2020, chemicals are used and produced in a way that minimizes ill effects on human health and the environment. Heads of state backed the goal in 2002. It has helped to implement national bans on the use of lead in paint in developing countries including Uruguay and Nepal.

Critics will say that the latest pledge is weak — and they are correct. It rejects specific actions to combat the problem, as had been proposed by the governments of Peru and Uruguay and by the International Society of Doctors for the Environment. There are no commitments to a network of scientists and experts to research and share knowledge, or to improve national bio-monitoring. And there are no new legal demands on drug firms to clean up their manufacturing processes.

Although drug companies say that they maintain good environmental practices, research shows that drug manufacture is a significant source of pharmaceutical pollution. For example, unpublished data from the US Geological Survey show that concentrations of certain drugs are up to five times higher in the effluents of wastewater-treatment plants that serve drug-manufacturing facilities compared with those that do not.

The powerful pharmaceutical and water industries, which fear expensive measures to help to address the problem, have already demonstrated their muscle. Through aggressive lobbying, they managed to derail European efforts to impose legal environmental limits on two drugs in 2012.

The ICCM agreement should help to change things. With the world's eyes now on this issue, industry groups and lobbyists will find it harder to bend initiatives in their favour. There could be an early test of the resolution: European policymakers plan to publish a strategy to tackle drug pollution in the region's waterways by the end of the year. ■

Optimistic outlook

In difficult times, Turkey is investing in a clutch of new scientific research centres.

The airy, architecturally striking building that is the brand-new Izmir Biomedicine and Genome Center (iBG) could have been anywhere in Europe. But when Turkey's science minister, Fikri İşik, turned up to speak at its inaugural ceremony last month (see page 171), surrounded by an ostentatious swarm of dark-clad security guards, the cultural differences were apparent.

Turkey, with its toe-hold on the European continent and a land-mass stretching nearly 2,000 kilometres to borders with Syria, Iraq, Iran, Georgia and Armenia, is familiar with difference. Mediterranean cities such as Istanbul and Izmir are westernized, but eastern cities are conservatively Islamic and the southeast is plagued by violence rooted in cross-border Kurdish separatist movements.

Politicians everywhere view the country as a potential bridge between the West and the war-torn Middle East, but some Turks fear that renewed Kurdish conflicts could degenerate into civil war. They fear also that the national election in three weeks will see president Recep Tayyip Erdoğan change the constitution to give himself still more power.

Can science thrive in this environment? Erdoğan has blurred the constitutional separation of state and religion. Under his regime, scientists have witnessed state-condoned rejection of Darwinism

and imprisonment of academics on trumped-up terrorism charges.

There are some positive signs, however. Turkey's negotiations with the European Union for membership stalled after troops violently dispersed political protestors in Istanbul's Taksim Square in 2013, yet the country continues to align its science policies with those of the EU. Accordingly, last year it passed two laws to improve and expand the research environment in strategic areas.

One law creates a slew of institutes across the country, some of which will provide regional services such as genome sequencing to ill-equipped universities. Others will be national research centres which, like the iBG, will aspire to carry out internationally competitive research and successfully compete to host major EU research facilities. Along with the advantage of secure funding, the national research centres will operate under new rules. They will be relatively free to manage their operations and budgets — a sign that the government recognizes that it should not micromanage research if it wants it to thrive.

The other law creates a Turkish National Institutes of Health, which will comprise 6 institutes, with the creation of 400 jobs in science.

All of these new centres must be allowed to develop free from political interference — scientists are particularly concerned that the government will seek close control over the health institutes.

Researchers in other Middle Eastern countries often find it simpler to collaborate with Turkish scientists than with westerners — travel is cheaper and usually visa-free. An improved scientific environment in Turkey may serve as the desired bridge, creating an intellectual network that can continue to converse, whatever the political tensions. Science can, in its limited way, be a force for peace. ■

➔ **NATURE.COM**
To comment online,
click on Editorials at:
go.nature.com/xhunqv



Make vaccine coverage a key UN health indicator

Track progress towards universal care using a wide-reaching intervention that all countries can readily measure, says Seth Berkley.

At the United Nations meeting in New York late last month, attendees started to refer to the new Sustainable Development Goals by a different name. The aims morphed into the Global Goals for sustainable development, or just Global Goals.

Whatever we call them, if the goals are to achieve what they set out to, the next few weeks will be crucial. At the end of this month, a UN expert group will meet to try to agree on how to measure progress — and success or failure.

Each of the 17 goals is made up of several targets — 169 in all. Global Goal 3, for example — to “ensure healthy lives and promote well-being for all at all ages” — includes a target to achieve universal health coverage (UHC). UHC is something that the World Health Organization has been pushing for since 2005, asking all countries to provide comprehensive health care for all citizens at an affordable cost.

The UN is exploring having each of these 169 targets judged against two ‘indicators’. But what can best indicate UHC? Unlike the Millennium Development Goals (MDGs) that preceded them, the Global Goals focus on both rich and poor countries. ‘Universal’ really must mean everyone.

One way to indicate progress towards UHC is to measure access to health interventions. But which treatments should we choose? Shine the spotlight on one and another is cast into the shadows. And how important is it for everyone to have access to the same treatments anyway? A child with type 1 diabetes growing up in Kansas clearly does not need the same access to mosquito nets as a child living in Somalia. And should we judge the health of the Somali child on the basis of their access to blood-glucose monitoring?

Given the challenge of trying to capture this complexity in a single measure, the UN is exploring having an indicator for UHC that is broken down into sub-indicators, which it calls tracers. Possible tracers include access to treatments for tuberculosis, hypertension and diabetes, as well as access to antiretroviral therapy and preventative measures for neglected tropical diseases. Others include improved sanitation, having a skilled attendant present during births, provision of insecticide-treated bed nets and access to full childhood immunization. In some countries, the list could extend to mental-health provision, treatment for cataracts, palliative care and other interventions.

At first glance, the list looks balanced. It reflects a good cross-section of disease burden, and each tracer can be monitored with relative ease using existing data sources such as health records or ones that can be readily set up, including household surveys. But does the list ensure the true health of a population?

Even if all countries made all these

interventions available, it would not necessarily mean that people were healthier. The fact that someone is in need of care suggests that they are not healthy, possibly because the system has in some way failed to prevent an illness.

With so many Global Goal targets — the eight MDGs had just 21 — there has been pressure on the UN to reduce the number of indicators. For UHC, one indicator is likely to be concerned with ‘affordability’, meaning that it is possible that all the chosen interventions, including those mentioned above, will be bundled into a single indicator.

This is a difficult problem. Even the common definition of ‘health’ as a state free from injury or disease is disputed by some. So it is no surprise that measuring health is fraught with problems. In trying to encompass this complexity, the UN risks creating an indicator that merely measures service coverage of a few selected therapeutic interventions.

Universal coverage is a means towards better health, but is not an end in itself. We should not be measuring health by access to treatments such as nicotine replacement therapy and lung surgery. Instead, we should be looking at tobacco control and other measures aimed at reducing smoking uptake in the first place.

A true indicator of UHC should be an intervention that every country can readily measure, that speaks to equitable access and quality, and that will reliably ensure the health of a population. Immunization is such an indicator. (Some data are missing, but all countries have agreed to work towards measuring vaccination rates.)

That is why some voices, including that of my organization, Gavi, the Vaccine Alliance, are calling for the Global Goals framework to make full childhood immunization a separate ambitious indicator of UHC in its own right.

More than 30 vaccine doses are administered globally every second. No other health intervention reaches so many people, or is capable of preventing such a diverse range of public-health concerns — from virulent infectious diseases such as measles, to cervical and liver cancer. And at the same time, it helps to identify worrying trends in rich countries — such as the drop in immunizations in parts of California to levels on a par with South Sudan, which has led to outbreaks in recent years.

If immunization is not made a separate indicator, then the UN should make clear that some of the tracers on its long list — including immunization — carry more weight than others. After all, as the old adage goes, when it comes to health, an ounce of prevention is worth a pound of cure. ■

Seth Berkley is chief executive of Gavi, the Vaccine Alliance, in Geneva, Switzerland.
e-mail: sberkley@gavi.org

**NO OTHER
HEALTH
INTERVENTION
REACHES
SO MANY
PEOPLE.**

➔ **NATURE.COM**
Discuss this article
online at:
go.nature.com/yeqaxu

RESEARCH HIGHLIGHTS

Selections from the
scientific literature

NEURODEGENERATION

Virus linked to neuron death

Viruses could be partly to blame for a neurodegenerative disease.

Avindra Nath at the National Institute of Neurological Disorders and Stroke in Bethesda, Maryland, and his colleagues studied post-mortem brain tissue from 11 people who had amyotrophic lateral sclerosis (ALS; also known as motor neuron disease), which leads to muscle weakness and paralysis. They found that neurons expressed a key protein from the human endogenous retrovirus-K, the DNA of which has been incorporated into human genomes over millions of years of evolution.

When the team introduced the gene encoding this protein into cultured human neurons, the cells decreased in number and retracted their neurites — the projections that connect to other cells. Mice expressing this protein showed a loss of muscle-controlling neurons and had muscle dysfunction.

Preventing this virus from being activated could slow the course of ALS, the authors say. *Sci. Transl. Med.* 7, 307ra153 (2015)

MOLECULAR BIOLOGY

DNA clusters help yeast in hard times

Starving yeast cells reorganize their chromosomes into dense clusters in a way that might slow the ageing process.

Angela Taddei from the Curie Institute in Paris used cell imaging and molecular-genetics techniques to visualize the 3D organization of chromosomes inside cells of baker's yeast (*Saccharomyces*

cerevisiae). The team found that starving yeast arrange their chromosomes so that their telomeres — long stretches of DNA at the ends of chromosomes that shorten with ageing — are tightly packed together in the centre of the cell's nucleus. The rearrangement is triggered by free radicals produced by the cell as it gradually exhausts the available food. Mutant strains that cannot produce these 'hyperclusters' do not survive starvation as well as their cluster-producing counterparts.

Packing telomeres together may prevent their degradation,

allowing dormant yeast cells to survive temporary food shortages.

Genome Biol. 16, 206 (2015)

GEOLOGY

Bigger volcanic blasts after impact

Volcanoes in India began spewing more magma around the time an asteroid hit Earth 66 million years ago. These eruptions and the asteroid impact both contributed to the mass extinction that killed off the dinosaurs.

A team led by Paul Renne at the Berkeley

Geochronology Center in California used argon isotopes to date volcanic rocks from India's Deccan Traps. They conclude that within 50,000 years of the impact, the volcanoes in this area began to pour out more lava with each eruption, even though the frequency of eruptions decreased.

Seismic waves generated by the impact, which happened off what is now Mexico's Yucatan peninsula, could have altered the geology of the Deccan Traps by expanding the size of magma chambers, for example. *Science* 350, 76–78 (2015)



VALERI YURKO

ECOLOGY

Animals thrive at Chernobyl

Wildlife populations seem to be increasing near the Chernobyl nuclear-disaster site, which people abandoned after a reactor explosion in 1986.

Jim Smith at the University of Portsmouth, UK, and his colleagues found that the Belarus sector of the exclusion zone around the devastated power plant had abundances of elk, deer and wild boar that were similar to those in four uncontaminated nature reserves in Belarus. Wolf numbers were more than seven times

higher around Chernobyl than in the other reserves. The team also found no correlation between contamination levels near the reactor site and the number of animal tracks.

The findings contradict previous studies suggesting that radiation around Chernobyl is harmful to wildlife populations, and show the resilience of large mammals to chronic radiation exposure, say the authors.

Curr. Biol. <http://dx.doi.org/10.1016/j.cub.2015.08.017> (2015)

AGRICULTURE

Wild field edges keep yields up

Setting aside some farmland as wildlife habitat might not reduce crop yields.

Richard Pywell at the Natural Environment Research Council's Centre for Ecology and Hydrology in Wallingford, UK, and his team studied 56 fields at a farm growing wheat, oilseed rape and field beans over 6 years. Along field edges, 0%, 3% or 8% of the total cropped area was set aside as habitat for birds, pollinators and other wildlife. None of the crop yields in the three experiments decreased, despite the difference in crop area. In fields without any habitat set aside, yields at the edges were poor, whereas in fields with habitat margins, the wildlife seemed to boost yields by increasing the productivity per unit area.

For beans, the yield was 35% higher in the fields where the most land was set aside.

Proc. R. Soc. B 282, 20151740 (2015)

DEVELOPMENTAL BIOLOGY

Limb and phallus share gene circuits

Limbed animals use the same genetic elements to regulate the development of their limbs and genitalia.

Douglas Menke and his colleagues at the University of Georgia in Athens studied the genomes and embryos of mice, *Anolis* lizards and snakes (pictured). They found that

in mice and lizards, many genetic regulatory elements are similarly active in the development of the limbs and the phallus. They also saw this pattern of activity in the external genitalia of the snake embryo (pictured, arrow).

Even though snakes lost their limbs during evolution, they probably retained the relevant genetic elements because of their importance for phallus development.

Dev. Cell <http://doi.org/743> (2015)

HYDROLOGY

Volcanoes change river flow

Tiny particles that are ejected into the atmosphere by volcanic eruptions can change the water cycle enough to alter the amount of water in nearby rivers for several years.

Carley Iles and Gabriele Hegerl at the University of Edinburgh, UK, looked at records of streamflow for 50 major rivers after volcanic eruptions dating back to 1883. Following an eruption, the amount of water flowing through the Amazon, Nile and other rivers in many tropical areas dropped for up to three years. In other areas, such as southwestern North America, streamflow increased.

Volcanic particles in the air reflect sunlight back into space, which cools the surface below, shifting patterns of evaporation and precipitation.

Water managers may need to plan for such interruptions, because of the importance of rivers for water supplies.

Nature Geosci. <http://dx.doi.org/10.1038/ngeo2545> (2015)

CANCER

How infection can cause leukaemia

Infection can trigger leukaemia in genetically susceptible mice, suggesting an environmental cause for the most common type of childhood cancer.

Children with precursor B-cell acute lymphoblastic leukaemia often have

SOCIAL SELECTION

Popular topics on social media

Campaign name draws criticism

Scientists generally laud attempts to get young people interested in careers in science, technology, engineering and mathematics (STEM). But an initiative to encourage girls to study science — launched by EDF Energy, a London-based power company — has met much scepticism and ridicule online, mostly because of the campaign's name: 'Pretty Curious' (go.nature.com/irijls). Michelle Kline, an anthropologist at Arizona State University in Tempe, tweeted: "#Prettycurious would be a good name for a dress line that uses science prints. But not for #womeninscience

where science comes first." The company responded to one critic by tweeting: "we deliberately chose the word 'pretty' to tackle the stereotype head on, #STEM careers should be accessible to all."

➔ **NATURE.COM**
For more on popular papers:
go.nature.com/xsntrc

mutations in the *PAX5* gene, which is involved in immune-cell development, but the mutations alone do not cause the disease. To see whether infection might be the trigger, Arndt Borkhardt at the University of Düsseldorf in Germany, Isidro Sanchez-Garcia at the University of Salamanca in Spain and their colleagues exposed mice with *Pax5* mutations to common pathogens. The mice developed cancer, whereas *Pax5* mutant mice kept in a sterile environment did not.

By sequencing tumour DNA from the diseased mice, the team found extra mutations — probably caused by infection — in genes encoding signalling proteins that help to regulate cell growth. Treating mice with molecules that inhibit these proteins lowered the number of cancer cells, suggesting a new avenue for treatment.

Cancer Discov. <http://doi.org/73j> (2015)

PLANT SCIENCE

Dung-like seeds dupe dung beetles

The seeds of a South African plant (pictured, top) trick dung beetles into dispersing

them by mimicking the appearance and odour of antelope faeces (pictured, bottom).

Jeremy Midgley at the University of Cape Town in South Africa and his colleagues observed dung beetles (*Epirinus flagellatus*) rolling seeds from the plant *Ceratocaryum argenteum* and burying them underground.

The authors then placed 195 seeds at 31 separate spots, and returned later to find that nearly half of the seeds had been dispersed.

Chemical analysis showed that the seeds emit volatile molecules similar to those in antelope dung.

The beetles had not nibbled on the seeds or deposited any eggs on them, suggesting that the insects are not aware of the deception until after they have planted the seeds. The beetles receive no apparent reward from this activity, the authors report.

Nature Plants <http://dx.doi.org/10.1038/nplants.2015.141> (2015)

➔ **NATURE.COM**
For the latest research published by Nature visit:
www.nature.com/latestresearch



NATURE PUBLISHING GROUP

CARLOS R. INFANTE

SEVEN DAYS

The news in brief

AWARDS

Nobel prizes

Three researchers who developed treatments for parasitic infections won the 2015 Nobel Prize in Physiology or Medicine. William Campbell and Satoshi Ōmura discovered a class of compounds called avermectins, which kill parasitic roundworms that cause infections such as river blindness; Youyou Tu developed the antimalarial drug artemisinin (see page 174). The physics prize was awarded to Takaaki Kajita and Arthur McDonald for their discovery of neutrino oscillations (see page 175). *Nature* went to press before the chemistry prize was awarded, but full details will be made available at go.nature.com/xkfab1.

Carbon XPRIZE

A new US\$20-million prize for technologies that can convert waste carbon dioxide from power plants into useful products was unveiled by the XPRIZE group on 29 September. The non-profit organization, based in Culver City, California, said that it hoped to award two \$7.5-million 'grand prizes' in March 2020. 'Milestone' prizes totalling \$5 million

NUMBER CRUNCH

31%

The percentage of World Heritage Sites — 70 of 229 — that are threatened by extractive industries such as oil and gas drilling, according to the World Wildlife Fund in a report on 1 October.

Source: World Wildl. Fund (2015)



CLOCKWISE FROM TOP LEFT: INSIGHTS/JUG/GETTY; PETER ESSICK/AURORA; TAMARA THOMSEN, WISCONSIN HISTORICAL SOC.; MALCOLM CLARK/NIAA

Marine-protection bonanza

World leaders unveiled plans for new 'marine protected areas' at an oceans conference in Valparaíso, Chile, on 5 October. The Chilean government will establish a 631,368-square-kilometre reserve around Easter Island in the Pacific (top left), and the United States will create its first national marine sanctuaries in 15 years — one off the coast of Maryland

(top right) and another in Lake Michigan (bottom right). On 28 September, New Zealand's President John Key announced the creation of a 620,000-square-kilometre sanctuary around the country's Kermadec Islands, a region that hosts huge populations of seabirds and marine animals, as well as the world's longest chain of underwater volcanos (bottom left).

will be awarded to ten teams in 2017.

PEOPLE

Forgery sentence

A Danish court sentenced neuroscientist Milena Penkowa to 9 months in prison on 30 October after finding her guilty of forgery related to research misconduct, but suspended the sentence. The City Court of Copenhagen said that Penkowa faked documents relating to the number of rodents used in experiments for her doctoral thesis. The University of Copenhagen, where she worked until 2010, and the Danish Committee on Scientific Dishonesty had previously concluded that Penkowa had committed

research misconduct. See go.nature.com/adhblu for more.

NIH appointment

Cardiologist Michael Lauer has been appointed chief of the extramural-research office at the US National Institutes of Health (NIH), which administers funds awarded to non-NIH employees, the agency announced on 28 September. The office disburses more than 80% of the NIH's US\$30-billion budget in grants, and sets policy in areas such as research-misconduct regulation. Lauer has headed the cardiovascular-research unit at the National Heart, Lung, and Blood Institute since 2009, and has worked mainly in epidemiology and biostatistics.

FUNDING

Neuroscience boost

The Kavli Foundation and its university partners announced on 1 October that they will spend more than US\$100 million on neuroscience research, including setting up three new Kavli neuroscience institutes for basic research. The US universities that will host the centres — Johns Hopkins University in Baltimore, Maryland, the Rockefeller University in New York City and the University of California, San Francisco — will co-finance them. Kavli, based in Oxnard, California, will also increase funding at its four existing neuroscience

SIMON DAWSON/BLOOMBERG/GETTY institutes, including that at the Norwegian University of Science and Technology in Trondheim, whose scientists shared last year's Nobel Prize in Physiology or Medicine.

US budget passes

The US Congress approved a temporary budget that continues funding for federal science agencies until 11 December, avoiding a forced shutdown of work at the US National Institutes of Health, National Science Foundation and NASA. Had US politicians failed to reach an agreement, funding for federal scientists and other employees would have ended on 1 October — the start of the country's 2016 fiscal year. But the 30 September decision just "kicks the can down the road", says Michael Lubell, director of public affairs at the American Physical Society in Washington DC. See go.nature.com/rlihoeg for more.

POLICY

Climate costs

The governor of the Bank of England has warned that climate change could lead to economic "tragedy". Mark Carney (pictured) told an insurance-industry meeting in London on 29 September that climate change "will threaten financial resilience and longer-term prosperity" by increasing



damage from storms and other natural disasters, and could upset financial markets. Carney is one of the most senior financiers to have taken such a strong stance on climate issues, and his speech has attracted some negative comment in the financial press.

Refugee scientists

The European Commission launched 'Science4Refugees', an initiative to help link refugee scientists with job openings, on 5 October. The service allows refugees to submit CV information to a web portal containing postings for jobs and fellowships at European institutions. Places that are open to employing refugees and asylum seekers will be marked with a Science4Refugees label, says the commission. Candidates compete for positions on the same basis as other applicants, and they need to have already obtained visas and work permits. In the long term, the commission intends to add

mentoring, language and other training opportunities. See go.nature.com/kzth45 for more.

RESEARCH

NASA eyes Venus

NASA's next Discovery-class planetary-exploration missions are targeting Venus and asteroids, the US space agency announced on 30 September. After whittling down 27 proposals for its US\$500-million venture, the agency has chosen 5 potential missions. Each one will receive \$3 million to develop its plans before one or two are selected to fly. Among those vying for lift-off are an orbiter to map Venus's surface, a probe to investigate its atmosphere, a telescope to hunt for near-Earth objects, a visit to the asteroid Psyche and a trip to four asteroids near Jupiter. See go.nature.com/pw723q for more.

HEALTH

Hit HIV early

Treatment for HIV should be provided immediately for anyone who is infected with the virus, advises the World Health Organization (WHO) in guidelines released on 30 September. These replace previous recommendations to start taking drugs only when immune-cell levels drop below a certain value, and it

COMING UP

8–9 OCTOBER

Science ministers from the G7 nations meet in Berlin.

go.nature.com/ss9idk

12–16 OCTOBER

Jerusalem hosts the 66th International Astronautical Congress.

go.nature.com/gv243k

11–14 OCTOBER

The International Cytokine & Interferon Society holds its annual meeting in Bamberg, Germany.

www.cytokines2015.com

expands the number of people who are eligible for treatment from roughly 28 million to 37 million worldwide. The WHO also calls for preventive drugs to be given to all people who are at substantial risk of HIV, rather than just to men who have sex with men. The guidance is based on evidence reported in July that earlier treatment helps both patients and public health (see *Nature* <http://doi.org/73w>; 2015).

BP settlement

The British oil giant BP will pay US\$20.8 billion to resolve civil lawsuits related to the Deepwater Horizon oil spill in 2010, the US Department of Justice announced on 5 October. Under the final settlement, reached with the US government and five states along the Gulf of Mexico, BP will pay \$5.5 billion in fines under the Clean Water Act and \$8.1 billion for natural-resource damages. Much of the money will be spent on coastal restoration projects. The company will also pay \$4.9 billion for economic impacts on the states and around \$2 billion in other payments.

➔ NATURE.COM

For daily news updates see:

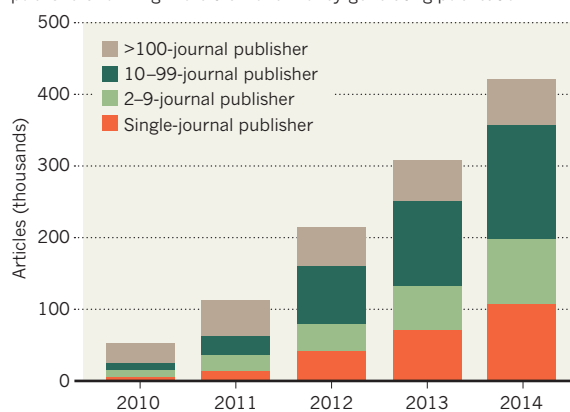
www.nature.com/news

TREND WATCH

"Predatory" open-access journals with "highly questionable" marketing and peer review are publishing more and more papers, finds a 1 October study (C. Shen and B.-C. Björk *BMC Med.* **13**, 230; 2015). These journals charge a fee for articles that undergo little or no editing or review; in 2010, they published 53,000 papers, rising to 420,000 in 2014. Authors paid an average of US\$178 per article. The team used titles from 'Beall's list' — which some claim includes legitimate publications (see *Nature* **495**, 433–435; 2013).

PREDATORS ON THE RISE

Dubious journals produce thousands of articles per year, with many publishers running more than one money-generating publication.



NEWS IN FOCUS

PHYSICS CERN tests small-scale plasma accelerator **p.173**

NOBEL PRIZE Malaria drug wins China its first Nobel **p.174**

SPACE NASA plans trips to Trojans and Psyche **p.176**



MATHEMATICS The proof so complex that no one can work out if it's right **p.178**

TUMAY BERKIN/ZUMA PRESS/CORBIS



A volatile political situation is intensifying the challenge of building a world-class research centre in Turkey.

BASIC RESEARCH

Turkish biomed hub spurs hope amid political strife

Centre in Izmir swims against the tide to produce world-class fundamental science.

BY ALISON ABBOTT

Rife with political tensions and notoriously unfriendly to basic research, Turkey may seem an unusual location to forge a hub of world-class biomedicine. But the Izmir Biomedicine and Genome Center (iBG), inaugurated last month in the ancient coastal city, has ambitions to be just that.

Based at the Dokuz Eylül University, the

iBG aspires to bridge a wide geographic gap in high-impact basic research that stretches from Europe to beyond India. At the same time, the centre intends to appease the Turkish government, which is keen for a return on its investments in science: at the inauguration ceremony for the iBG on 10 September, science minister Fikri Işık stated that he expected the centre to “start making money”.

If the institute succeeds, it will be thanks

in part to the deftness of its director, Mehmet Öztürk, at balancing the two demands. In particular, he has found ways to exploit the government's interest in applied science and funnel some of it towards basic research.

“It is a huge experiment,” says Hermann Bujard, a member of the iBG's scientific advisory board and a molecular biologist at the University of Heidelberg in Germany. “If it takes off, it could be a crystallization point ▶

that helps change the situation for science in Turkey — and also serve as a bridge to Middle Eastern countries.”

The experiment is taking place in a politically volatile environment. An election on 1 November — the second this year — will decide whether the ruling Justice and Development Party (AKP) can secure the absolute majority it needs to change the constitution in a way that some fear would move the country close to dictatorship. And a ceasefire between Turkey and Kurdish separatists broke down a few months ago, leading to outbreaks of violence that have killed hundreds.

At the same time, the government is trying to do something about its low investment in research as part of a general strategy to align its policies to those of the European Union, which it is still negotiating to join. Its research spending is creeping up, but, in relation to gross domestic product, the nation still spends less than half of the EU average. And the vast majority of that is devoted to applied science.

NEGOTIATING TACTICS

In early 2013, Dokuz Eylül University offered Öztürk a job leading a new institute that would be devoted to translational research — such as turning biomedical discoveries made elsewhere into drugs or diagnostics. Öztürk hesitated. A molecular biologist who has spent much of his career abroad, he believed that the centre needed to do fundamental research, too. “Basic research is generally considered a waste of time and money in Turkey,” he says. “But translation can’t be maintained long-term in an intellectual vacuum.”

It took him six months to convince the university’s rector, but that September, Öztürk took up the post under the agreement that the institute would incorporate both types of research. Even so, he knew that he was taking a risk. “I understood that to keep basic research going, we’d also have to give those investing in our institute what they really want: applications and services.”

He has since filled around 20 of the planned 32 positions for principal investigators,



Mehmet Öztürk has devised ways to create funding for basic research.

drawing from both within and outside Turkey. All are in the early stages of their careers and bring national and international grant money with them — an estimated 88 million lira (US\$29 million) for 2016.

Still, to ensure the centre’s long-term sustainability, Öztürk needs a more reliable source of support than competitive grants.

“I understood that to keep basic research going, we’d also have to give those investing in our research what they really want.”

low salaries.

Help with both problems may be close. Although not yet a member of the EU, Turkey would like to compete to host some parts of the international research infrastructures being coordinated by the European Commission. To increase the country’s chances of success, the government last year passed a law in which it agreed to pay for the construction, operation and staffing costs of research centres with the sophistication to host multimillion-euro collaborative projects.

The law also releases the centres from the

notorious bureaucracy that governs Turkish public institutions, giving them more freedom to manage themselves and set their own salaries.

In the next few months, the government will select 10–15 centres to fund; the iBG is considered a strong candidate.

Öztürk is also exploiting the government’s desire to develop an industry around production of the expensive biological medicines that it currently imports to treat cancer and other diseases. Since 2013, the national research agency TÜBİTAK has put out calls to fund research to develop such ‘biosimilars’. The iBG plans to offer a service to industries who want to try their hand at this, including facilities that can offer

manufacturing at international clinical-safety standards and access to university hospital beds for clinical trials. Eventually, this division would be spun off into a company owned by the iBG and the profits would subsidize basic research, says Öztürk.

Moreover, all iBG principal investigators will be asked to participate in translational projects in parallel with their own basic research, a policy that could bring in yet more funding. TÜBİTAK offers large innovation grants to support such translational work that include a generous 40% for overheads, which, Öztürk says, could be used to support basic research at the iBG.

Tim Hunt, a molecular biologist at the Crick Institute in London and a guest speaker at the inauguration of the iBG, says that the creation of the institute is “an astonishing thing in a country not known for its science”. He adds that it is a “fantastic opportunity” for talented Turkish scientists to return to their country — and to a well-equipped and well-funded lab.

But the precarious politics of Turkey are never completely absent. At the iBG inauguration, Dokuz Eylül University president Mehmet Füzün paid respect to tens of Turkish soldiers who had died in the previous days in roadside bombings related to the separatist struggles. Acute fears of civil war have faded since then, but scientists are always aware that politics could jeopardize the iBG dream. ■ **SEE EDITORIAL P.164**

HANI ALQATBI



Island boulders reveal ancient mega-tsunamis from Cape Verde volcano
go.nature.com/wnffku

- Gains in Antarctic ice might offset losses go.nature.com/elzzfz
- Archaeologists ousted by ISIS return to Iraqi cave go.nature.com/byqvif
- Strong placebo response thwarts painkiller trials go.nature.com/nbushq



The impenetrable proof; toggling REM sleep; and the latest from the Rosetta mission
nature.com/nature/podcast

PHYSICS

CERN to test revolutionary mini-accelerator

Plasma wakefield machines aim to reach high energies without huge gains in size.

BY ELIZABETH GIBNEY

The home of the Large Hadron Collider (LHC), the world's largest particle accelerator, is getting a new machine — and this time, the whole point is to keep it small.

On 18 September, the council that governs CERN, Europe's premier particle-physics laboratory, near Geneva, Switzerland, approved a boost in funding for a planned experiment called the Advanced Wakefield Experiment, or AWAKE. Due to switch on next year, AWAKE will accelerate particles by 'surfing' them on waves of electric charge created in a plasma, or ionized gas. It is a method that could allow future accelerators to probe matter and the forces of nature at ever-higher energies, without the usual accompanying increase in the instruments' size and therefore cost.

Although plans are afoot to build bigger machines once the LHC reaches the end of its life in the 2030s (see go.nature.com/a9sm2m), many fear that accelerator size is nearing its limit and that such proposals may simply prove too expensive to implement.

"When you look at cost estimates for these machines and the scale of machines, you understand that maybe a new breakthrough regime is needed," says Nick Walker, an accelerator physicist at DESY, Germany's high-energy-physics laboratory in Hamburg.

Conventional colliders, such as the 27-kilometre-long LHC, use electric fields to move charged particles through a tunnel; the fields switch from positive to negative at a frequency that means the particles are constantly nudged forward, gaining energy with each push. But such colliders use metal-walled cavities that spark if the electric field is too strong. As a result, the only way to further increase the particles' speed, and therefore energy, is to lengthen the tunnel.

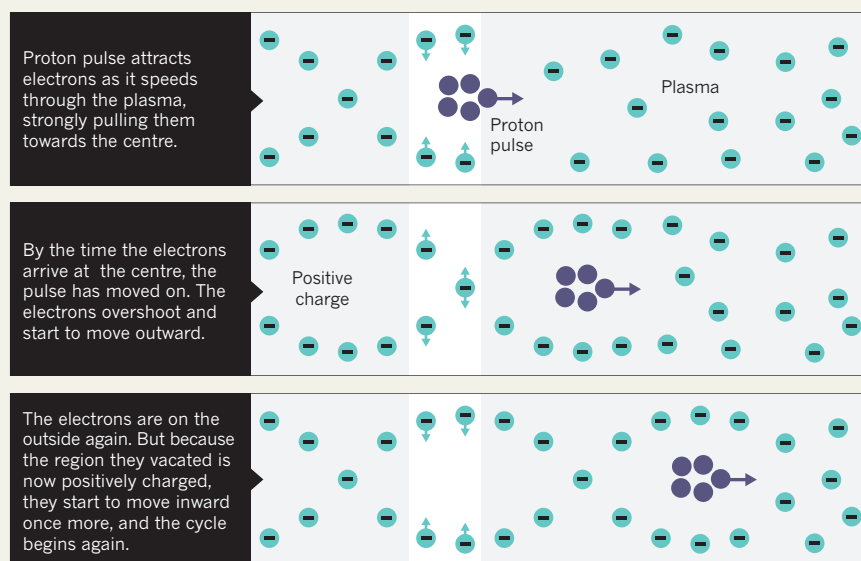
Plasma wakefield accelerators, which were first proposed in the 1970s, are designed to break this cycle, says physicist Allen Caldwell at the Max Planck Institute for Physics in Munich, Germany, who will lead the AWAKE experiment. They send a pulse of charged particles or laser light through a plasma, which sets electrons and positively charged ions oscillating in its wake. The resulting regions of alternating negative and positive charge form waves that accelerate further charged particles. Injected

WAKEFIELD ACCELERATION

The AWAKE experiment at CERN will test whether pulses of protons can be used to turn a plasma-filled machine into a particle accelerator.

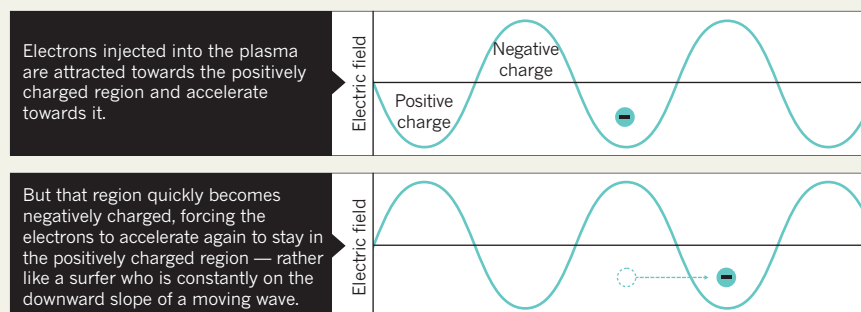
MAKING WAVES

A pulse of protons injected into an ionized gas, or plasma, sets electrons bobbing in its wake, creating regions that constantly cycle between being positively and negatively charged.



SURF'S UP

The cycling from positive to negative charge creates a wave of electron density that can be used to accelerate injected electrons.



at just the right time, these particles effectively surf the waves (see 'Wakefield acceleration'). Crucially, as the electric fields are much stronger than those in a conventional collider, the acceleration can be as much as 1,000 times greater over the same distance.

Such accelerators exist in prototype at several facilities around the world, but AWAKE

will be the first time that CERN has experimented with the technology. "CERN is the world's high-energy physics lab right now, and the fact that it has decided this is an important field to get involved in is a bit of validation for this community," says Mark Hogan, an accelerator physicist at the SLAC National Accelerator Laboratory in Menlo Park, California.

Different groups have different ways of setting the plasma oscillating: Hogan's team at SLAC uses pulses of electrons, for example. AWAKE will be the first to use pulses of protons, which have some big advantages.

Because protons have greater mass than electrons, each proton pulse penetrates further into the plasma, setting up a longer series of charged regions, which in turn provides greater acceleration per pulse. A proton machine is also compatible with the LHC, which accelerates and collides protons.

For now, AWAKE will use the proton bunches that feed the LHC to test whether protons can generate the electric fields necessary to accelerate particles in plasma.

The latest investment from CERN — worth 2.6 million Swiss francs (US\$2.7 million), from the total of 21.4 million Swiss francs so far committed to the experiment — is intended to allow AWAKE to test the concept before the end of 2018, when CERN is scheduled to shut down its accelerators for an upgrade. Success will depend

“The fact that CERN has decided this is an important field to get involved in is a bit of validation for this community.”

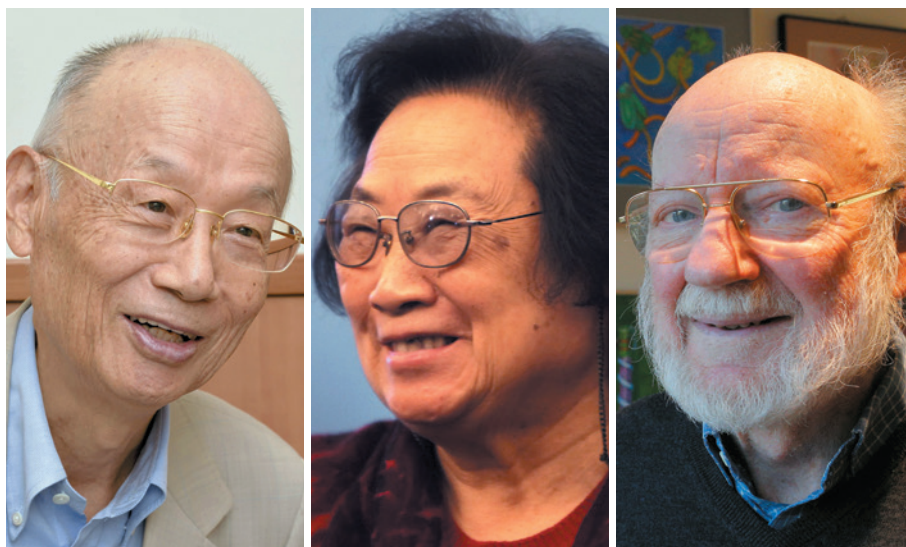
on whether these proton bunches, which are long relative to what is needed to create plasma waves, can be efficiently chopped up into short pulses.

Eventually, it might be possible

to inject the much-higher-energy protons that have been accelerated by the LHC into a plasma wakefield machine for further acceleration. Hogan estimates that a machine just a few kilometres long could produce electrons with 6 times the energy of those that would be produced by the next planned conventional accelerator, the 31-kilometre-long International Linear Collider.

Despite such promise, plasma accelerators are decades from practical use because, to do better than existing accelerators, they must also match them in efficiency — supplying focused, accelerated particles at high rates as well as high energies, says Walker. Still, he adds, “right now, this is the only thing I see that might work”.

The technology might also be useful elsewhere. Wakefield-accelerated electrons could drive X-ray free-electron lasers, which probe matter using powerful bursts of light that are short enough to capture the motions of molecules. These are currently kilometres long — but using wakefield technology might allow them to fit into labs or hospital basements. “I think this is more realistic as a potential application,” says Walker, “and I would say a mandatory first step before the plunge into trying to achieve high-energy-physics experiments.” ■



Satoshi Ōmura, Youyou Tu and William C. Campbell share the Nobel Prize in Physiology or Medicine.

MEDICINE

China celebrates first Nobel

Pharmacologist shares prize for work on parasitic infections.

BY EWEN CALLAWAY & DAVID CYRANOSKI

For the first time, a researcher based in China has won the ultimate status symbol in science — a Nobel prize.

Pharmacologist Youyou Tu, who led a Beijing team that discovered the key malaria drug artemisinin in the late 1960s and 1970s, was awarded the 2015 Nobel Prize in Physiology or Medicine on 5 October. Two microbiologists, William C. Campbell at Drew University in Madison, New Jersey, and Satoshi Ōmura at Kitasato University in Japan, shared the award for their development — also in the 1970s — of therapies against parasitic roundworms.

“This certainly is fantastic news for China. We expect more to come in the future,” says Wei Yang, president of the nation’s main research-funding agency, the National Natural Science Foundation of China. Lan Xue, an innovation-studies specialist at Tsinghua University in Beijing, says that he was inundated with messages about the prize. “People will be celebrating, but I hope they also take a sober look, because there are lots of things to learn from this award,” he says.

Young scientists in China today are told to go overseas to do good research and to churn out publications in internationally recognized journals, Xue notes. Yet Tu has never worked outside China, and has not racked up major

publications. “Tu doesn’t fit into any of the trends today, and yet she gets the Nobel because of the originality of her work. It couldn’t have been a better choice in terms of the lessons it offers Chinese scientists,” Xue says.

MALARIA BREAKTHROUGH

Tu’s prizewinning research, at the China Academy of Chinese Medical Sciences in Beijing, originated from a government push in 1967 to discover new therapies for malaria. At the time, the main treatments — chloroquine and the older quinine — were proving increasingly ineffective. Tu and her team screened more than 2,000 Chinese herbal remedies to search for drugs with antimalarial activity. An extract from the wormwood plant *Artemisia annua* proved especially effective, and by 1972, the researchers had isolated chemically pure artemisinin.

“I’m very happy about this. She totally deserves it,” says Yi Rao, a neuroscientist at Peking University in Beijing who has researched the discovery of artemisinin. But there has been some controversy over credit for the discovery, Rao points out, so Tu has never won a major award in China. She has not been elected to either of China’s major academies — neither the Chinese Academy of Sciences nor the Chinese Academy of Engineering.

“Though other people were involved, Tu

THE YOMURI SHIMBUN VIA AP IMAGES, REUTERS/STRINGER, REUTERS/BRIAN SNYDER

was clearly the undisputed leader,” says Rao. “But she’s never been given fair recognition within China.”

Artemisinin has “saved possibly millions of lives”, says Stephen Ward at the Liverpool School of Tropical Medicine, UK. And the work of Campbell and Ōmura, who together discovered a class of compounds known as avermectins that kill parasitic roundworms that cause infections such as river blindness and lymphatic filariasis, has protected millions from disease, he adds.

Working in Japan, Ōmura isolated strains of a group of soil bacteria called *Streptomyces* that

were known to have antimicrobial properties. In 1974, he pulled a promising organism out of soil near a golf course and sent it, along with others, to a team led by Campbell at the Merck Institute for Therapeutic Research in Rahway, New Jersey. (Ōmura’s institute had signed a research partnership with Merck in 1973.)

Campbell’s team isolated avermectins from the bacterial cultures and tweaked the structure of one of the most promising compounds to develop it into a drug — ivermectin. In 1987, Merck announced that it would donate the drug to anyone who needed it for treatment of onchocerciasis (also known as river blindness).

A decade later, the firm began giving away the drug to treat lymphatic filariasis. Each year, Merck gives away some 270 million treatments of the drug, according to the Mectizan Donation Program in Decatur, Georgia.

Ward notes that the Nobel this year highlights the global acceptance of the importance of parasitic infections and of neglected tropical diseases in general. “It may refocus us on the idea that the immense diversity of products out there in the natural world is a great starting point for drug discovery,” he says. ■

Additional reporting by Alison Abbott.

NOBEL PRIZE

Neutrino flip wins physics prize

Physicists share Nobel for solving puzzle about the subatomic particles’ changing identities.

BY ELIZABETH GIBNEY &
DAVIDE CASTELVECCHI

Two researchers who helped to demonstrate that neutrinos oscillate between types, or ‘flavours’, as they travel — which proved that the elusive particles have mass — have won this year’s Nobel Prize in Physics.

Takaaki Kajita at the University of Tokyo and Arthur McDonald at Queen’s University in Kingston, Canada, share the prize for their discoveries with teams at two deep, underground neutrino detectors — Kajita at the Super-Kamiokande neutrino detector in Hida, Japan, and McDonald at the Sudbury Neutrino Observatory in Canada.

The standard model of particle physics — the current best explanation of the Universe’s particles and forces — struggles to explain why neutrinos have mass. So the two teams’ discoveries, in 1998 and 2001, spurred a wave of new experiments seeking to pin down the neutrino’s properties. “Other than the Higgs boson, I’d say this is the biggest discovery in particle physics in the last 30 years,” says Daniel Hooper, a theoretical physicist at the University of Chicago in Illinois.

Neutrinos are more abundant than any other particle in the Universe except for the photon: each second, billions of them stream through every square centimetre of Earth. But they interact so weakly with other matter that remarkably little is known about them.

The first hint that neutrinos were stranger than expected came in the 1960s. Physicists knew that neutrinos come in three flavours: electron, muon and tau, names that relate to the sister particle they are produced with. But an experiment at the Homestake gold mine in South Dakota threw up a mystery: it detected



Takaaki Kajita and Arthur McDonald share the 2015 Nobel Prize in Physics.

fewer electron-type neutrinos streaming from the Sun than theorists had predicted. (Alongside Masatoshi Koshihara of the University of Tokyo, Raymond Davis, who led the Homestake experiment, later shared half of the 2002 Nobel Prize in Physics for developing techniques to detect such neutrinos from space.)

Kajita’s group began unravelling this conundrum in 1998, when it reported that neutrinos might change flavours as they travel. Muon neutrinos created in collisions between cosmic rays and Earth’s atmosphere seemed to disappear on their way to the Super-Kamiokande detector, a steel tank filled with pure water located in a zinc mine.

Conclusively proving this, however, meant not just spotting ‘disappearing’ neutrinos, but showing that they had turned into other flavours. The Sudbury team, using a tank of water in a nickel mine more than 2,000 metres

beneath Earth’s surface, announced in 2001 that neutrinos oscillated between flavours as they travelled from the Sun to Earth.

The discovery has profound implications. Rather than the three neutrino flavours having no mass, or indeed any fixed masses, physicists now reason that neutrinos must be made from mixtures — or quantum superpositions — of three different mass states, which change in proportion as the particles travel. Pinning down the neutrino properties and their antimatter counterpart, antineutrinos, could lead to an understanding of physics beyond the standard model, says André Rubbia, a neutrino physicist at the Swiss Federal Institute of Technology in Zurich.

“We believe that differences in the way neutrinos and antineutrinos oscillate, for example, is the best possible explanation we have for why the Universe is today dominated by matter and not antimatter,” says Rubbia. ■

ENVIRONMENT

India unveils climate pledge

Country seeks big cuts in carbon intensity and greater reliance on clean energy.

BY T. V. PADMA

India says that it will produce 40% of its energy from sources other than fossil fuels by 2030, and will reduce the intensity of its carbon dioxide emissions by roughly one-third.

The country's highly anticipated announcement on 2 October comes ahead of United Nations talks in Paris this December, at which nations hope to reach an updated agreement to fight climate change.

India is the third-largest emitter of greenhouse gases, and it is the last major economy to announce its climate commitment ahead of the Paris meeting. But it is also a nation where 300 million people still lack access to electricity, and its per-capita greenhouse-gas emissions are well below the global average.

"India is not part of the problem" of global warming, says environment minister Prakash Javadekar. "But we want to be part of the solution." He calls the country's plan "comprehensive, ambitious and progressive".

The pledge eschews an overall cap on CO₂ emissions, in an effort to protect vulnerable sectors of India's economy and society. Instead, India says that it will reduce its carbon intensity — emissions per unit of gross domestic product — by 33–35% in 2030, compared with the 2005 level. Javadekar estimates

that meeting this goal will prevent 3.59 billion tonnes of CO₂ emissions.

The country will also aim to generate 40% of its electricity from renewable or low-carbon sources by 2030, with technology-transfer and financial assistance from the Green Climate Fund, an organization headquartered in

"India is not part of the problem — but we want to be part of the solution."

Songdo, South Korea, that was formed to help developing nations to address climate change.

Carrying out the entire plan will cost at least US\$2.5 trillion, the government says, with some of that money coming as international aid.

"India traditionally has taken a very hard line in the negotiations, and done its best to avoid assuming obligations," says Elliot Diringer, executive vice-president of the Center for Climate and Energy Solutions in Arlington, Virginia. "This reflects a very encouraging shift in attitude toward an acceptance that all major economies share a responsibility to address this challenge."

MIXED REACTION

Others were more critical of the new plan. Navroz Dubash, a senior fellow at the Centre for Policy Research (CPR) in New Delhi, says that the carbon-intensity goal

is "conservative at best". It is well below the 45% intensity cut recommended in a draft 2015 report by the CPR and the International Institute for Applied Systems Analysis in Laxenburg, Austria.

And Dubash notes that the plan does not offer many details on policies for specific economic sectors. "We will need more transparency, monitoring and assessment down the line to see what the sectoral actions add up to, and whether they will help India avoid a lock-in into a high-carbon growth pathway," he says.

Shreekanth Gupta, an economist at India's Delhi School of Economics, approves of the pledge's emphasis on promoting economic growth and access to energy to reduce poverty. But Gupta would have liked a more radical approach to these issues, such as a cap-and-trade scheme patterned after the European Union's emissions-trading programme.

And Chandra Venkataraman, a chemical engineer at the Indian Institute of Technology in Mumbai, says that India has missed an opportunity to reduce its emissions of black carbon, a sooty pollutant that is produced by the incomplete burning of biomass and other fuels. Black carbon harms human health, and it has potent — although relatively short-lived — warming effects on the climate. ■

Additional reporting by Jeff Tollefson.

SPACE

NASA picks finalists

Venus and asteroids make shortlist of planetary missions.

BY ALEXANDRA WITZE

Venus and asteroids have emerged as top destinations for NASA's future planetary exploration. On 30 September, the agency announced a shortlist of five contenders for its US\$500-million Discovery class of missions.

Two of the five proposed missions would target Venus, which NASA has not visited in more than two decades. A radar orbiter would map the planet's cloud-enshrouded surface from above, while an atmospheric probe would descend directly through the layers of haze. "They're pretty exciting choices and

focus on a body that has not received much attention," says Steven Hauck, a planetary scientist at Case Western Reserve University in Cleveland, Ohio.

Asteroid mission concepts include a telescope to hunt for dangerous near-Earth objects; a visit to the peculiarly metal-rich asteroid Psyche; and a tour of four Trojan asteroids that orbit near Jupiter.

NASA will give each of the proposed missions \$3 million to develop their ideas further, and by September 2016, the agency will select one — or possibly two, if budgets permit — to eventually fly. (The space agency started with a list of 27 candidates.)

The selection capped months of anxious waiting for many US planetary scientists, who submitted their ideas in February. "It's been an amazing day," says Harold Levison, a planetary scientist at the Southwest Research Institute in Boulder, Colorado, who heads the Trojan asteroid proposal. "I got the call when I was driving to work," he says. "I pulled over."

In principle, the Discovery competition is open for ideas to visit any target in the Solar System other than Earth or the Sun.

Among the mission concepts that lost out were a spacecraft to whizz past erupting volcanoes on Jupiter's moon Io, and one to analyse plumes spewing from Saturn's moon Enceladus, which many consider a promising place for extraterrestrial life. Also left on the sidelines were several proposals to study comets, and three focusing on Mars's moons.

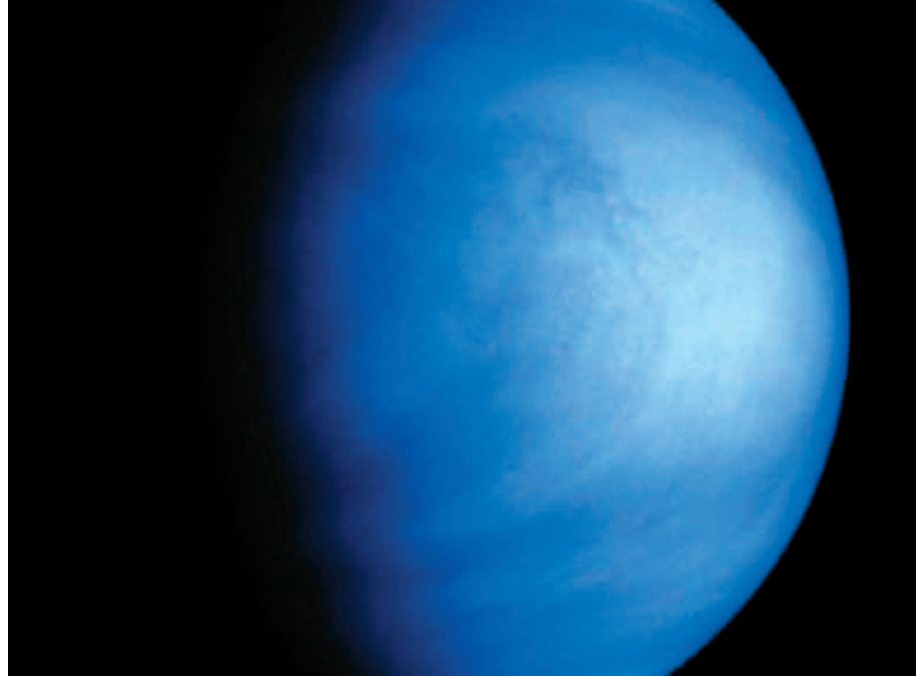
Women lead four of the five shortlisted missions. Suzanne Smrekar, of the Jet Propulsion Laboratory (JPL) in Pasadena, California, heads the VERITAS mission to

map Venus at higher resolution and in different radar frequencies than NASA's Magellan mission of the early 1990s. Lori Glaze, of NASA's Goddard Space Flight Center in Greenbelt, Maryland, is leading development of a probe that would descend through Venus's atmosphere over about an hour, making measurements along the way.

Farther out in the Solar System, planetary scientist Lindy Elkins-Tanton heads the push to visit the asteroid Psyche. It represents a primordial not-quite-planet whose outer rocky layers have been stripped away to expose its metallic heart. "Psyche is the only core that humankind can ever see," says Elkins-Tanton, of Arizona State University in Tempe. "We've visited gassy things and rocky things and icy things, but we've never visited a metal object," she says. The Psyche mission would launch in 2020 for a 2026 arrival.

Levison's 'Lucy' — named after the famous early-human fossil — would fly past a main-belt asteroid on its way to visit the Trojans. These poorly understood space rocks share an orbit with Jupiter but may have originated farther out in the Solar System. "There's a huge diversity in this population, and that diversity is telling us about the evolution of the Solar System," he says. Lucy would launch in 2021 and reach the end of its mission in 2032.

The asteroid mission NEOCam (for



The blue tints in this image of Venus reveal variations in the clouds that surround the planet.

Near-Earth Object Camera) would use an infrared telescope to hunt for small and faint but potentially hazardous asteroids. Led by Amy Mainzer of JPL, it has been through the Discovery selection process twice before; NASA rejected the proposal in 2006 but gave Mainzer money in 2010 to develop the telescope's infrared detectors. "We really want to go find some asteroids and settle the question of whether one is heading our way," Mainzer says.

Although Discovery missions are supposed to launch every couple of years, the current candidates are the first to be selected since 2010. Ongoing Discovery missions include the Dawn spacecraft, which is orbiting the asteroid Ceres, and the Kepler telescope that searches for extrasolar planets. In March, NASA plans to launch its next Discovery spacecraft, InSight, which will place a seismometer on the surface of Mars. ■



The impenetrable proof

Shinichi Mochizuki claims to have solved one of the most important problems in mathematics. The trouble is, hardly anyone can work out whether he's right.

By Davide Castelvecchi

Sometime on the morning of 30 August 2012, Shinichi Mochizuki quietly posted four papers on his website.

The papers were huge — more than 500 pages in all — packed densely with symbols, and the culmination of more than a decade of solitary work. They also had the potential to be an academic bombshell. In them, Mochizuki claimed to have solved the *abc* conjecture, a 27-year-old problem in number theory that no other mathematician had even come close to solving. If his proof was correct, it would be one of the most astounding achievements of mathematics this century and would completely revolutionize the study of equations with whole numbers.

Mochizuki, however, did not make a fuss about his proof. The respected mathematician, who works at Kyoto University's Research Institute for Mathematical Sciences (RIMS) in Japan, did not even announce his work to peers around the world. He simply posted the papers, and waited for the world to find out.

Probably the first person to notice the papers was Akio Tamagawa, a colleague of Mochizuki's at RIMS. He, like other researchers, knew

that Mochizuki had been working on the conjecture for years and had been finalizing his work. That same day, Tamagawa e-mailed the news to one of his collaborators, number theorist Ivan Fesenko of the University of Nottingham, UK. Fesenko immediately downloaded the papers and started to read. But he soon became “bewildered”, he says. “It was impossible to understand them.”

Fesenko e-mailed some top experts in Mochizuki's field of arithmetic geometry, and word of the proof quickly spread. Within days, intense chatter began on mathematical blogs and online forums (see *Nature* <http://doi.org/725>; 2012). But for many researchers, early elation about the proof quickly turned to scepticism. Everyone — even those whose area of expertise was closest to Mochizuki's — was just as flummoxed by the papers as Fesenko had been. To complete the proof, Mochizuki had invented a new branch of his discipline, one that is astonishingly abstract even by the standards of pure maths. “Looking at it, you feel a bit like you might be reading a paper from the future, or from outer space,” number theorist Jordan Ellenberg, of the University of Wisconsin–Madison, wrote on his blog a few days after the paper appeared.

Three years on, Mochizuki's proof remains in mathematical limbo — neither debunked nor accepted by the wider community. Mochizuki has estimated that it would take an expert in arithmetic geometry some 500 hours to understand his work, and a maths graduate student about ten years. So far, only four mathematicians say that they have been able to read the entire proof.

Adding to the enigma is Mochizuki himself. He has so far lectured about his work only in Japan, in Japanese, and despite being fluent in English, he has declined invitations to talk about it elsewhere. He does not speak to journalists; several requests for an interview for this story went unanswered. Mochizuki has replied to e-mails from other mathematicians and been forthcoming to colleagues who have visited him, but his only public input has been sporadic posts on his website. In December 2014, he wrote that to understand his work, there was a “need for researchers to deactivate the thought patterns that they have installed in their brains and taken for granted for so many years”. To mathematician Lieven Le Bruyn of the University of Antwerp in the Netherlands, Mochizuki's attitude sounds defiant. “Is it just me,” he wrote on his blog earlier this year, “or is Mochizuki really sticking up his middle finger to the mathematical community?”

Now, that community is attempting to sort the situation out. In December, the first workshop on the proof outside of Asia will take

place in Oxford, UK. Mochizuki will not be there in person, but he is said to be willing to answer questions from the workshop through Skype. The organizers hope that the discussion will motivate more mathematicians to invest the time to familiarize themselves with his ideas — and potentially move the needle in Mochizuki's favour.

In his latest verification report, Mochizuki wrote that the status of his theory with respect to arithmetic geometry “constitutes a sort of faithful miniature model of the status of pure mathematics in human society”. The trouble that he faces in communicating his abstract work to his own discipline mirrors the challenge that mathematicians as a whole often face in communicating their craft to the wider world.

Primal importance

The *abc* conjecture refers to numerical expressions of the type $a + b = c$. The statement, which comes in several slightly different versions, concerns the prime numbers that divide each of the quantities a , b and c . Every whole number, or integer, can be expressed in an essentially unique way as a product of prime numbers — those that cannot be further factored out into smaller whole numbers: for example, $15 = 3 \times 5$ or $84 = 2 \times 2 \times 3 \times 7$. In principle, the prime factors of a and b have no connection to those of their sum, c . But the *abc* conjecture links them together. It presumes, roughly, that if a lot of small primes divide a and b then only a few, large ones divide c .

This possibility was first mentioned in 1985, in a rather off-hand remark about a particular class of equations by French mathematician Joseph Oesterlé during a talk in Germany. Sitting in the audience was David Masser, a fellow number theorist now at the University of Basel in Switzerland, who recognized the potential importance of the conjecture, and later publicized it in a more general form. It is now credited to both, and is often known as the Oesterlé–Masser conjecture.

A few years later, Noam Elkies, a mathematician at Harvard University in Cambridge, Massachusetts, realized that the *abc* conjecture, if true, would have profound implications for the study of equations concerning whole numbers — also known as Diophantine equations after Diophantus, the ancient-Greek mathematician who first studied them.

Elkies found that a proof of the *abc* conjecture would solve a huge collection of famous and unsolved Diophantine equations in one stroke. That is because it would put explicit bounds on the size of the solutions. For example, *abc* might show that all the solutions to an equation must be smaller than 100. To find those solutions, all one would have to do would be to plug in every number from 0 to 99 and calculate which ones work. Without *abc*, by contrast, there would be infinitely many numbers to plug in.

Elkies's work meant that the *abc* conjecture could supersede the most important breakthrough in the history of Diophantine equations: confirmation of a conjecture formulated in 1922 by the US mathematician Louis Mordell, which said that the vast majority of Diophantine equations either have no solutions or have a finite number of them. That conjecture was proved in 1983 by German mathematician Gerd Faltings, who was then 28 and within three years would win a Fields Medal, the most coveted mathematics award, for the work. But if *abc* is true, you don't just know how many solutions there are, Faltings says, “you can list them all”.

Soon after Faltings solved the Mordell conjecture, he started teaching at Princeton University in New Jersey — and before long, his path crossed with that of Mochizuki.

Born in 1969 in Tokyo, Mochizuki spent his formative years in the United States, where his family moved when he was a child. He attended

an exclusive high school in New Hampshire, and his precocious talent earned him an undergraduate spot in Princeton's mathematics department when he was barely 16. He quickly became legend for his original thinking, and moved directly into a PhD.

People who know Mochizuki describe him as a creature of habit with an almost supernatural ability to concentrate. “Ever since he was a student, he just gets up and works,” says Minhyong Kim, a mathematician at the University of Oxford, UK, who has known Mochizuki since his Princeton days. After attending a seminar or colloquium, researchers and students would often go out together for a beer — but not Mochizuki, Kim recalls. “He's not introverted by nature, but he's so much focused on his mathematics.”

Faltings was Mochizuki's adviser for his senior thesis and for his doctoral one, and he could see that Mochizuki stood out. “It was clear that he was one of the brighter ones,” he says. But being a Faltings student couldn't have been easy. “Faltings was at the top of the intimidation ladder,” recalls Kim. He would pounce on mistakes, and when talking to him, even eminent mathematicians could often be heard nervously clearing their throats.

Faltings's research had an outsized influence on many young number theorists at universities along the US eastern seaboard.

His area of expertise was algebraic geometry, which since the 1950s had been transformed into a highly abstract and theoretical field by Alexander Grothendieck — often described as the greatest mathematician of the twentieth century. “Compared to Grothendieck,” says Kim, “Faltings didn't have as much patience for philosophizing.” His style of maths required “a lot of abstract background knowledge — but also tended to have as a goal very concrete problems. Mochizuki's work on *abc* does exactly this”.

Single-track mind

After his PhD, Mochizuki spent two years at Harvard and then in 1994 moved back to his native Japan, aged 25, to a position at RIMS. Although he had lived for years in the United States, “he was in some ways uncomfortable with American culture”, Kim says. And, he adds, growing up in a different country may have compounded the feeling of isolation that comes from being a mathematically gifted child. “I think he did suffer a little bit.”

Mochizuki flourished at RIMS, which does not require its faculty members to teach undergraduate classes. “He was able to work on his own for 20 years without too much external disturbance,” Fesenko says. In 1996, he boosted his international reputation when he solved a conjecture that had been stated by Grothendieck; and in 1998, he gave an invited talk at the International Congress of Mathematicians in Berlin — the equivalent, in this community, of an induction to a hall of fame.

But even as Mochizuki earned respect, he was moving away from the mainstream. His work was reaching higher levels of abstraction and he was writing papers that were increasingly impenetrable to his peers. In the early 2000s he stopped venturing to international meetings, and colleagues say that he rarely leaves the Kyoto prefecture any more. “It requires a special kind of devotion to be able to focus over a period of many years without having collaborators,” says number theorist Brian Conrad of Stanford University in California.

Mochizuki did keep in touch with fellow number theorists, who knew that he was ultimately aiming for *abc*. He had next to no competition: most other mathematicians had steered clear of the problem, deeming it intractable. By early

“Looking at it, you feel a bit like you might be reading a paper from the future.”

► NATURE.COM

To hear a podcast on Shinichi Mochizuki's proof, visit: go.nature.com/v6rfy7

2012, rumours were flying that Mochizuki was getting close to a proof. Then came the August news: he had posted his papers online.

The next month, Fesenko became the first person from outside Japan to talk to Mochizuki about the work he had quietly unveiled. Fesenko was already due to visit Tamagawa, so he went to see Mochizuki too. The two met on a Saturday in Mochizuki's office, a spacious room offering a view of nearby Mount Daimonji and with neatly arranged books and papers. It is "the tidiest office of any mathematician I've ever seen in my life", Fesenko says. As the two mathematicians sat in leather armchairs, Fesenko peppered Mochizuki with questions about his work and what might happen next.

Fesenko says that he warned Mochizuki against speaking to the press about his proof. He was mindful of the experience of another mathematician: the Russian topologist Grigori Perelman, who shot to fame in 2003 after solving the century-old Poincaré conjecture (see *Nature* 427, 388; 2004) and then retreated and became increasingly estranged from friends, colleagues and the outside world. Fesenko knew Perelman, and thinks that his behaviour was a result of excessive media attention. But Fesenko soon saw that the two mathematicians' personalities could not have been more different. Whereas Perelman was known for his awkward social skills (and for letting his fingernails grow unchecked), Mochizuki is universally described as articulate and friendly — if intensely private about his life outside of work.

Normally after a major proof is announced, mathematicians read the work — which is typically a few pages long — and can understand the general strategy. Occasionally, proofs are longer and more complex, and years may then pass for leading specialists to fully vet it and reach a consensus that it is correct. Perelman's work on the Poincaré conjecture became accepted in this way. Even in the case of Grothendieck's highly abstract work, experts were able to relate most of his new ideas to mathematical objects they were familiar with. Only once the dust has settled does a journal typically publish the proof.

But almost everyone who tackled Mochizuki's proof found themselves floored. Some were bemused by the sweeping — almost messianic — language with which Mochizuki described some of his new theoretical instructions: he even called the field that he had created 'inter-universal geometry'. "Generally, mathematicians are very humble, not claiming that what they are doing is a revolution of the whole Universe," says Oesterlé, at the Pierre and Marie Curie University in Paris, who made little headway in checking the proof.

The reason is that Mochizuki's work is so far removed from anything that had gone before. He is attempting to reform mathematics from the ground up, starting from its foundations in the theory of sets (familiar to many as Venn diagrams). And most mathematicians have been reluctant to invest the time necessary to understand the work because they see no clear reward: it is not obvious how the theoretical machinery that Mochizuki has invented could be used to do calculations. "I tried to read some of them and then, at some stage, I gave up. I don't understand what he's doing," says Faltings.

Fesenko has studied Mochizuki's work in detail over the past year, visited him at RIMS again in the autumn of 2014 and says that he has now verified the proof. (The other three mathematicians who say they have corroborated it have also spent considerable time working alongside Mochizuki in Japan.) The overarching theme of inter-universal geometry, as Fesenko describes it, is that one must look at whole numbers in a different light — leaving addition aside and seeing the multiplication structure as something malleable and deformable. Standard multiplication would then be just one particular case of a family of structures, just as a circle is a special case of an ellipse.

Fesenko says that Mochizuki compares himself to the mathematical giant Grothendieck — and it is no immodest claim. "We had mathematics before Mochizuki's work — and now we have mathematics after Mochizuki's work," Fesenko says.

But so far, the few who have understood the work have struggled to explain it to anyone else. "Everybody who I'm aware of who's come close to this stuff is quite reasonable, but afterwards they become incapable of communicating it," says one mathematician who did not want his name to be mentioned. The situation, he says, reminds him of the *Monty Python* skit about a writer who jots down the world's funniest joke. Anyone who reads it dies from laughing and can never relate it to anyone else.

And that, says Faltings, is a problem. "It's not enough if you have a good idea: you also have to be able to explain it to others." Faltings says that if Mochizuki wants his work to be accepted, then he should reach out more. "People have the right to be eccentric as much as they want to," he says. "If he doesn't want to travel, he has no obligation. If he wants recognition, he has to compromise."

Edge of reason

For Mochizuki, things could begin to turn around later this year, when the Clay Mathematics Institute will host the long-awaited workshop in Oxford. Leading figures in the field are expected to attend, including Faltings. Kim, who along with Fesenko is one of the organizers, says that a few days of lectures will not be enough to expose the entire theory. But, he says, "hopefully at the end of the workshop enough people will be convinced to put more of their effort into reading the proof".

Most mathematicians expect that it will take many more years to find some resolution. (Mochizuki has said that he has submitted his papers to a journal, where they are presumably still under review.) Eventually, researchers hope, someone will be willing not only to understand the work, but also to make it understandable to others — the problem is, few want to be that person.

Looking ahead, researchers think that it is unlikely that future open problems will be as complex and intractable. Ellenberg points out that theorems are generally simple to state in new mathematical fields, and the proofs are quite short.

The question now is whether Mochizuki's proof will edge towards acceptance, as Perelman's did, or find a different fate. Some researchers see a cautionary tale in that of Louis de Branges, a well-established mathematician at Purdue University in West Lafayette, Indiana. In 2004, de Branges released a purported solution to the Riemann hypothesis, which many consider the most important open problem in maths. But mathematicians have remained sceptical of that claim; many say that they are turned off by his unconventional theories and his idiosyncratic style of writing, and the proof has slipped out of sight.

For Mochizuki's work, "it's not all or nothing", Ellenberg says. Even if the proof of the *abc* conjecture does not work out, his methods and ideas could still slowly percolate through the mathematical community, and researchers might find them useful for other purposes. "I do think, based on my knowledge of Mochizuki, that the likelihood that there's interesting or important math in those documents is pretty high," Ellenberg says.

But there is still a risk that it could go the other way, he adds. "I think it would be pretty bad if we just forgot about it. It would be sad." ■

Davide Castelvecchi is a reporter for *Nature* in London.

"I tried to read some of them and then, at some stage, I gave up."



FOOLING OURSELVES

**HUMANS ARE REMARKABLY GOOD AT SELF-DECEPTION.
BUT GROWING CONCERN ABOUT REPRODUCIBILITY IS DRIVING MANY
RESEARCHERS TO SEEK WAYS TO FIGHT THEIR OWN WORST INSTINCTS.**

BY REGINA NUZZO

In 2013, five years after he co-authored a paper showing that Democratic candidates in the United States could get more votes by moving slightly to the right on economic policy¹, Andrew Gelman, a statistician at Columbia University in New York City, was chagrined to learn of an error in the data analysis. In trying to replicate the work, an undergraduate student named Yang Yang Hu had discovered that Gelman had got the sign wrong on one of the variables.

Gelman immediately published a three-sentence correction, declaring that everything in the paper's crucial section should be considered wrong until proved otherwise.

ILLUSTRATION BY DALE EDWIN MURRAY

Reflecting today on how it happened, Gelman traces his error back to the natural fallibility of the human brain: “The results seemed perfectly reasonable,” he says. “Lots of times with these kinds of coding errors you get results that are just ridiculous. So you know something’s got to be wrong and you go back and search until you find the problem. If nothing seems wrong, it’s easier to miss it.”

This is the big problem in science that no one is talking about: even an honest person is a master of self-deception. Our brains evolved long ago on the African savannah, where jumping to plausible conclusions about the location of ripe fruit or the presence of a predator was a matter of survival. But a smart strategy for evading lions does not necessarily translate well to a modern laboratory, where tenure may be riding on the analysis of terabytes of multidimensional data. In today’s environment, our talent for jumping to conclusions makes it all too easy to find false patterns in randomness, to ignore alternative explanations for a result or to accept ‘reasonable’ outcomes without question — that is, to ceaselessly lead ourselves astray without realizing it.

Failure to understand our own biases has helped to create a crisis of confidence about the reproducibility of published results, says statistician John Ioannidis, co-director of the Meta-Research Innovation Center at Stanford University in Palo Alto, California. The issue goes well beyond cases of fraud. Earlier this year, a large project that attempted to replicate 100 psychology studies managed to reproduce only slightly more than one-third². In 2012, researchers at biotechnology firm Amgen in Thousand Oaks, California, reported that they could replicate only 6 out of 53 landmark studies in oncology and haematology³. And in 2009, Ioannidis and his colleagues described how they had been able to fully reproduce only 2 out of 18 microarray-based gene-expression studies⁴.

Although it is impossible to document how often researchers fool themselves in data analysis, says Ioannidis, findings of irreproducibility beg for an explanation. The study of 100 psychology papers is a case in point: if one assumes that the vast majority of the original researchers were honest and diligent, then a large proportion of the problems can be explained only by unconscious biases. “This is a great time for research on research,” he says. “The massive growth of science allows for a massive number of results, and a massive number of errors and biases to study. So there’s good reason to hope we can find better ways to deal with these problems.”

“When crises like this issue of reproducibility come along, it’s a good opportunity to advance our scientific tools,” says Robert MacCoun, a social scientist at Stanford. That has happened before, when scientists in the mid-twentieth century realized that experimenters and subjects often unconsciously changed their behaviour to match expectations. From that insight, the double-blind standard was born.

“People forget that when we talk about the scientific method, we don’t mean a finished product,” says Saul Perlmutter, an astrophysicist at the University of California, Berkeley. “Science is an ongoing race between our inventing ways to fool ourselves, and our inventing ways to avoid fooling ourselves.” So researchers are trying a variety of creative ways to debias data analysis — strategies that involve collaborating with academic rivals, getting papers accepted before the study has even been started and working with strategically faked data.

“SCIENCE IS AN ONGOING RACE BETWEEN OUR INVENTING WAYS TO FOOL OURSELVES, AND OUR INVENTING WAYS TO AVOID FOOLING OURSELVES.”

THE PROBLEM

Although the human brain and its cognitive biases have been the same for as long as we have been doing science, some important things have changed, says psychologist Brian Nosek, executive director of

the non-profit Center for Open Science in Charlottesville, Virginia, which works to increase the transparency and reproducibility of scientific research. Today’s academic environment is more competitive than ever. There is an emphasis on piling up publications with statistically significant results — that is, with data relationships in which a commonly used measure of statistical certainty, the *p*-value, is 0.05 or less. “As a researcher, I’m not trying to produce misleading results,” says Nosek. “But I do have a stake in the outcome.” And that gives the mind excellent motivation to find what it is primed to find.

Another reason for concern about cognitive bias is the advent of staggeringly large multivariate data sets, often harbouring only a faint signal in a sea of random noise. Statistical methods have barely caught up with such data, and our brain’s methods are even worse, says Keith Baggerly, a statistician at the University of Texas MD Anderson Cancer Center in Houston. As he told a conference on challenges in bioinformatics last September in Research Triangle Park, North Carolina, “Our intuition when we start looking at 50, or hundreds of, variables sucks.”

Andrew King, a management specialist at Dartmouth College in Hanover, New Hampshire, says that the widespread use of point-and-click data-analysis software has made it easy for researchers to sift through massive data sets without fully understanding the methods, and to find small *p*-values that may not actually mean anything. “I believe we are in the steroids era of social science,” he says. “I’ve been guilty of using some of these performance-enhancing practices myself. My sense is that most researchers have fallen at least once.”

Just as in competitive sport, says Hal Pashler, a psychologist at the University of California, San Diego, this can set up a vicious circle of chasing increasingly better results. When a few studies in behavioural neuroscience started reporting improbably strong correlations of 0.85, Pashler says, researchers who had more moderate (and plausible) results started to worry: “Gee, I just got a 0.4, so maybe I’m not really doing this very well.”

HYPOTHESIS MYOPIA

One trap that awaits during the early stages of research is what might be called hypothesis myopia: investigators fixate on collecting evidence to support just one hypothesis; neglect to look for evidence against it; and fail to consider other explanations. “People tend to ask questions that give ‘yes’ answers if their favoured hypothesis is true,” says Jonathan Baron, a psychologist at the University of Pennsylvania in Philadelphia.

For example, says Baron, studies have tried to show how disgust influences moral condemnation, “by putting the subject in a messy room, or a room with ‘fart spray’ in the air”. The participants are then asked to judge how to respond to moral transgressions; if those who have been exposed to clutter or smells favour harsher punishments, researchers declare their ‘disgust hypothesis’ to be supported⁵. But they have not considered competing explanations, he says, and so they ignore the possibility that participants are lashing out owing to anger at their foul treatment, not simply disgust. By focusing on one hypothesis, researchers might be missing the real story entirely.

Courtrooms face a similar problem. In 1999, a woman in Britain called Sally Clark was found guilty of murdering two of her sons, who had died suddenly as babies. A factor in her conviction was the presentation of statistical evidence that the chances of two children in the same family dying of sudden infant death syndrome (SIDS) were only 1 in 73 million — a figure widely interpreted as fairly damning. Yet considering just one hypothesis leaves out an important part of the story. “The jury needs to weigh up two competing explanations for the babies’ deaths: SIDS or murder,” wrote statistician Peter Green on behalf of the Royal Statistical Society in 2002 (see go.nature.com/ochsja). “The

fact that two deaths by SIDS is quite unlikely is, taken alone, of little value. Two deaths by murder may well be even more unlikely. What matters is the relative likelihood of the deaths under each explanation, not just how unlikely they are under one explanation.” Mathematician Ray Hill of the University of Salford, UK, later estimated⁶ that a double SIDS death would occur in roughly 1 out of 297,000 families, whereas two children would be murdered by a parent in roughly 1 out of 2.7 million families — a likelihood ratio of 9 to 1 against murder. In 2003, Clark’s conviction was overturned on the basis of new evidence. The Attorney General for England and Wales went on to release two other women who had been convicted of murdering their children on similar statistical grounds.

THE TEXAS SHARPSHOOTER

A cognitive trap that awaits during data analysis is illustrated by the fable of the Texas sharpshooter: an inept marksman who fires a random pattern of bullets at the side of a barn, draws a target around the biggest clump of bullet holes, and points proudly at his success.

His bullseye is obviously laughable — but the fallacy is not so obvious to gamblers who believe in a ‘hot hand’ when they have a streak of wins, or to people who see supernatural significance when a lottery draw comes up as all odd numbers.

Nor is it always obvious to researchers. “You just get some encouragement from the data and then think, well, this is the path to go down,” says Pashler. “You don’t realize you had 27 different options and you picked the one that gave you the most agreeable or interesting results, and now you’re engaged in something that’s not at all an unbiased representation of the data.”

Psychologist Uri Simonsohn at the University of Pennsylvania, gives an explicit nod to this naivety in his definition of ‘*p*-hacking’: “Exploiting — perhaps unconsciously — researcher degrees of freedom until $p < 0.05$.” In 2012, a study of more than 2,000 US psychologists⁷ suggested how common *p*-hacking is. Half had selectively reported only studies that ‘worked’, 58% had peeked at the results and then decided whether to collect more data, 43% had decided to throw out data only after checking its impact on the *p*-value and 35% had reported unexpected findings as having been predicted from the start, a practice that psychologist Norbert Kerr of Michigan State University in East Lansing has called HARKing, or hypothesizing after results are known. Not only did the researchers admit to these *p*-hacking practices, but they defended them.

This May, a journalist described how he had teamed up with a German documentary filmmaker and demonstrated that creative *p*-hacking, carried out over one “beer-fueled” weekend, could be used to ‘prove’ that eating chocolate leads to weight loss, reduced cholesterol levels and improved well-being (see go.nature.com/blkpke). They gathered 18 different measurements — including weight, blood protein levels and sleep quality — on 15 people, a handful of whom had eaten some extra chocolate for a few weeks. With that many comparisons, the odds were better than 50–50 that at least one of them would look statistically significant just by chance. As it turns out, three of them did — and the team cherry-picked only those to report.

ASYMMETRIC ATTENTION

The data-checking phase holds another trap: asymmetric attention to detail. Sometimes known as disconfirmation bias, this happens when we give expected results a relatively free pass, but we rigorously check non-intuitive results. “When the data don’t seem to match previous estimates, you think, ‘Oh, boy! Did I make a mistake?’” MacCoun

**“WHEN THE DATA DON’T
SEEM TO MATCH PREVIOUS
ESTIMATES, YOU THINK,
‘OH, BOY! DID I MAKE A
MISTAKE?’”**

says. “We don’t realize that probably we would have needed corrections in the other situation as well.”

The evidence suggests that scientists are more prone to this than one would think. A 2004 study⁸ observed the discussions of researchers from 3 leading molecular-biology laboratories as they worked through 165 different lab experiments. In 88% of cases in which results did not align with expectations, the scientists blamed the inconsistencies on how the experiments were conducted, rather than on their own theories. Consistent results, by contrast, were given little to no scrutiny.

In 2011, an analysis of over 250 psychology papers found⁹ that more than 1 in 10 of the *p*-values was incorrect — and that when the errors were big enough to change the statistical significance of the result, more than 90% of the mistakes were in favour of the researchers’ expectations, making a non-significant finding significant.

JUST-SO STORYTELLING

As data-analysis results are being compiled and interpreted, researchers often fall prey to just-so storytelling — a fallacy named after the Rudyard Kipling tales that give whimsical explanations for things such as how the leopard got its spots. The problem is that post-hoc stories can be concocted to justify anything and everything — and so end up truly explaining nothing. Baggerly says that he has seen such stories in genetics studies, when an analysis implicates a huge number of genes in a particular trait or outcome. “It’s akin to a Rorschach test,” he said at the bioinformatics conference. Researchers will find a story, he says, “whether it’s there or not. The problem is that occasionally it ain’t real.”

Another temptation is to rationalize why results should have come up a certain way but did not — what might be called JARKing, or justifying after results are known. Matthew Hankins, a statistician at King’s College London, has collected more than 500 creative phrases that researchers use to convince readers that their non-significant results are worthy of attention (see go.nature.com/pwctoq). These include “flirting with conventional levels of significance ($p > 0.1$)”, “on the very fringes of significance ($p = 0.099$)” and “not absolutely significant but very probably so ($p > 0.05$)”.

THE SOLUTIONS

In every one of these traps, cognitive biases are hitting the accelerator of science: the process of spotting potentially important scientific relationships. Countering those biases comes down to strengthening the ‘brake’: the ability to slow down, be sceptical of findings and eliminate false positives and dead ends.

One solution that is piquing interest revives an old tradition: explicitly considering competing hypotheses, and if possible working to develop experiments that can distinguish between them. This approach, called strong inference¹⁰, attacks hypothesis myopia head on. Furthermore, when scientists make themselves explicitly list alternative explanations for their observations, they can reduce their tendency to tell just-so stories.

In 2013, researchers reported¹¹ using strong-inference techniques in a study of what attracts female túngara frogs (*Engystomops pustulosus*) during mating calls. The existing data could be explained equally well by two competing theories — one in which females have a preset neural template for mating calls, and another in which they flexibly combine auditory cues and visual signals such as the appearance of the males’ vocal sacs. So the researchers developed an experiment for which the

two theories had opposing predictions. The results showed that females can use multi-sensory cues to judge attractiveness.

TRANSPARENCY

Another solution that has been gaining traction is open science. Under this philosophy, researchers share their methods, data, computer code and results in central repositories, such as the Center for Open Science's Open Science Framework, where they can choose to make various parts of the project subject to outside scrutiny. Normally, explains Nosek, "I have enormous flexibility in how I analyse my data and what I choose to report. This creates a conflict of interest. The only way to avoid this is for me to tie my hands in advance. Precommitment to my analysis and reporting plan mitigates the influence of these cognitive biases."

An even more radical extension of this idea is the introduction of registered reports: publications in which scientists present their research plans for peer review before they even do the experiment. If the plan is approved, the researchers get an 'in-principle' guarantee of publication, no matter how strong or weak the results turn out to be. This should reduce the unconscious temptation to warp the data analysis, says Pashler. At the same time, he adds, it should keep peer reviewers from discounting a study's results or complaining after results are known. "People are evaluating methods without knowing whether they're going to find the results congenial or not," he says. "It should create a much higher level of honesty among referees." More than 20 journals are offering or plan to offer some format of registered reports.

TEAM OF RIVALS

When it comes to replications and controversial topics, a good debiasing approach is to bypass the typical academic back-and-forth and instead invite your academic rivals to work with you. An adversarial collaboration has many advantages over a conventional one, says Daniel Kahneman, a psychologist at Princeton University in New Jersey. "You need to assume you're not going to change anyone's mind completely," he says. "But you can turn that into an interesting argument and intelligent conversation that people can listen to and evaluate." With competing hypotheses and theories in play, he says, the rivals will quickly spot flaws such as hypothesis myopia, asymmetric attention or just-so storytelling, and cancel them out with similar slants favouring the other side.

Psychologist Eric-Jan Wagenmakers of the University of Amsterdam has engaged in this sort of proponent-sceptic collaboration, when he teamed up with another group in an attempt¹² to replicate its research suggesting that horizontal eye movements help people to retrieve events from their memory. It is often difficult to get researchers whose original work is under scrutiny to agree to this kind of adversarial collaboration, he says. The invitation is "about as attractive as putting one's head on a guillotine — there is everything to lose and not much to gain". But the group that he worked with was eager to get to the truth, he says. In the end, the results were not replicated. The sceptics remained sceptical, and the proponents were not convinced by a single failure to replicate. Yet this was no stalemate. "Although our adversarial collaboration has not resolved the debate," the researchers wrote, "it has generated new testable ideas and has brought the two parties slightly closer." Wagenmakers suggests several ways in which this type of collaboration could be encouraged, including a prize for best adversarial collaboration, or special sections for such collaborations in top journals.

**"I'M NOT TRYING TO
PRODUCE MISLEADING
RESULTS — BUT I DO
HAVE A STAKE IN THE
OUTCOME."**

BLIND DATA ANALYSIS

One debiasing procedure has a solid history in physics but is little known in other fields: blind data analysis (see page 187). The idea is that researchers who do not know how close they are to desired results will be less likely to find what they are unconsciously looking for¹³.

One way to do this is to write a program that creates alternative data sets by, for example, adding random noise or a hidden offset, moving participants to different experimental groups or hiding demographic categories. Researchers handle the fake data set as usual — cleaning the data, handling outliers, running

analyses — while the computer faithfully applies all of their actions to the real data. They might even write up the results. But at no point do the researchers know whether their results are scientific treasures or detritus. Only at the end do they lift the blind and see their true results — after which, any further fiddling with the analysis would be obvious cheating.

Perlmutter used this method for his team's work on the Supernova Cosmology Project in the mid-2000s. He knew that the potential for the researchers to fool themselves was huge. They were using new techniques to replicate estimates of two crucial quantities in cosmology — the relative abundances of matter and of dark energy — which together reveal whether the Universe will expand forever or eventually collapse into a Big Crunch. So their data were shifted by an amount known only to the computer, leaving them with no idea what their findings implied until everyone agreed on the analyses and the blind could be safely lifted. After the big reveal, not only were the researchers pleased to confirm earlier findings of an expanding Universe¹⁴, Perlmutter says, but they could be more confident in their conclusions. "It's a lot more work in some sense, but I think it leaves you feeling much safer as you do your analysis," he says. He calls blind data analysis "intellectual hygiene, like washing your hands".

Data blinding particularly appeals to young researchers, Perlmutter says — not least because of the sense of suspense it gives. He tells the story of a recent graduate student who had spent two years under a data blind as she analysed pairs of supernova explosions. After a long group meeting, Perlmutter says, the student presented all her analyses and said that she was ready to unblind if everyone agreed.

"It was 6 o'clock in the evening and time for dinner," says Perlmutter. And everyone in the audience said, "If the result comes out wrong, it's going to be a very disappointing evening, and she's going to have to think really hard about what she's going to do with her PhD thesis. Maybe we should wait until morning."

"And we all looked at each other, and we said, 'Nah! Let's unblind now!' So we unblinded, and the results looked great, and we all cheered and applauded." ■ [SEE COMMENT P.187 & P.189](#)

Regina Nuzzo is a freelance writer in Washington DC.



NATURE.COM
For Nature's special
collection on
reproducibility, see:
go.nature.com/huhbyr

- Gelman, A. & Cai, C. J. *Ann. Appl. Stat.* **2**, 536–549 (2008).
- Open Science Collaboration. *Science* <http://dx.doi.org/10.1126/science.aac4716> (2015).
- Begley, C. G. & Ellis, L. M. *Nature* **483**, 531–533 (2012).
- Ioannidis, J. P. A. et al. *Nature Genet.* **41**, 149–155 (2009).
- Landy, J. F. & Goodwin, G. P. *Perspect. Psychol. Sci.* **10**, 518–536 (2015).
- Hill, R. *Paediatr. Perinatal Epidemiol.* **18**, 320–326 (2004).
- John, L. K., Loewenstein, G. & Prelec, D. *Psychol. Sci.* **23**, 524–532 (2012).
- Fugelsang, J. A., Stein, C. B., Green, A. E. & Dunbar, K. N. *Can. J. Exp. Psychol.* **58**, 86–95 (2004).
- Bakker, M. & Wicherts, J. M. *Behav. Res. Meth.* **43**, 666–68 (2011).
- Platt, J. R. *Science* **146**, 347–353 (1964).
- Taylor, R. C. & Ryan, M. J. *Science* **341**, 273–274 (2013).
- Matzke, D. et al. *J. Exp. Psychol. Gen.* **144**, e1–e15 (2015).
- MacCoun, R. & Perlmutter, S. in *Psychological Science Under Scrutiny: Recent Challenges and Proposed Solutions* (eds Lilienfeld, S. O. & Waldman, I.) (Wiley, in the press); Preprint available at <http://ssrn.com/abstract=2563337>
- Conley, A. et al. *Astrophys. J.* **644**, 1–20 (2006).

COMMENT

REPRODUCIBILITY Twenty-nine teams, one data set, one question, many answers **p.189**

EARTH Biography of Alfred Wegener, discoverer of continental drift **p.192**

FILM Ridley Scott delivers a rose-tinted take on the red planet **p.193**



OBITUARY Eric Davidson, systems-biology pioneer, remembered **p.196**

ILLUSTRATION BY DALE EDWIN MURRAY



Hide results to seek the truth

More fields should, like particle physics, adopt blind analysis to thwart bias, urge **Robert MacCoun** and **Saul Perlmutter**.

Decades ago, physicists including Richard Feynman noticed something worrying. New estimates of basic physical constants were often closer to published values than would be expected given standard errors of measurement¹. They realized that researchers were more likely to ‘confirm’ past results than refute them — results that did not conform to their expectation were more often systematically discarded or revised.

To minimize this problem, teams of particle physicists and cosmologists developed methods of blind analysis: temporarily and judiciously removing data labels and altering data values to fight bias and error². By the early 2000s, the technique had become widespread in areas of particle and nuclear physics. Since 2003, one of us (S.P.) has, with colleagues,

been using blind analysis for measurements of supernovae that serve as a ‘cosmic yardstick’ in studies of the unexpected acceleration of the Universe’s expansion³.

In several subfields of particle physics and cosmology, a new sort of analytical culture is forming: blind analysis is often considered the only way to trust many results. It is also being used in some clinical protocols (the term ‘triple-blinding’ sometimes refers to this⁴), and is increasingly used in forensic laboratories as well.

But the concept is hardly known in the biological, psychological and social

sciences. One of us (R.M.) has considerable experience conducting empirical research on legal and public-policy controversies in which concerns about bias are rampant (for example, drug legalization), but first encountered the concept when the two of us co-taught a transdisciplinary course at the University of California, Berkeley, on critical thinking and the role of science in democratic group decision-making. We came to recognize that the methods that physicists were using might improve trust and integrity in many sciences, including those with high-stakes analyses that are easily plagued by bias.

Many motivations distort what inferences we draw from data. These include the desire to support one’s theory, to refute one’s competitors, to be first to report a phenomenon, or simply to avoid publishing ‘odd’ ►



NATURE.COM
For Nature’s special collection on reproducibility, see: go.nature.com/huhbyr

► results. Such biases can be conscious or unconscious. They can occur irrespective of whether choices are motivated by the search for truth, by the good mentor's desire to help their student write a strong PhD thesis, or just by naked self-interest⁵.

We argue that blind analysis should be used more broadly in empirical research. Working blind while selecting data and developing and debugging analyses offers an important way to keep scientists from fooling themselves.

WHO KNOWS WHAT

Some forms of blinding are well known: for example, shielding both patients and clinicians from knowing who receives an experimental drug or a placebo (double-blinding), or removing names and affiliations from scientific manuscripts to keep peer reviewers from being swayed by authors' identities. But these practices apply to the collection and source of data, rather than the analysis.

Blind analysis ensures that all analytical decisions have been completed, and all programmes and procedures debugged, before relevant results are revealed to the experimenter. One investigator — or, more typically, a suitable computer program — methodically perturbs data values, data labels or both, often with several alternative versions of perturbation. The rest of the team then conducts as much analysis as possible 'in the dark'. Before unblinding, investigators should agree that they are sufficiently confident of their analysis to publish whatever the result turns out to be, without further rounds of debugging or rethinking. (There is no barrier to conducting extra analyses once data are unblinded, but doing so risks bias, so researchers should label such further analyses as 'post-blind'.)

There are many ways to do blind analysis. The computer need not (and probably will not) be blinded to data values; it is the display of results that masks information. Techniques must obscure meaningful results while showing enough of the data's structure to allow researchers to find and debug measurement artefacts, irrelevant variables, spurious correlates and other problems. For example, researchers who analyse clinical-trial results without knowing which patients received a placebo should still be able to identify implausible values.

The best methods for blinding depend on the properties of the data (for example, the type of statistical distribution, lower and upper bounds, whether values are discrete or continuous and whether cases were randomly assigned to experimental conditions or passively observed). Both data values and labels can be manipulated to develop a suitable

"Blinding analyses could be as simple as asking a colleague to scramble labels."

BLINDING STRATEGIES

| Technique examples | Perturbation | Potential application |
|---|---|---|
| Noising $\theta_{ij} = y_{ij} + n_{ij}$ or $\theta_{ij} = \beta_k + n_{ij}$ | Add a random number (from an appropriate statistical distribution) to data points or model parameters. | Testing which of several prevention messages is most effective in reducing smoking. |
| Biasing $\theta_{ij} = y_{ij} + b_j$ | Obscure differences in experimental conditions by adding a hidden value that is biased in a particular direction. | Estimating whether the costs of a controversial safety regulation exceed its benefits. |
| Cell scrambling $\theta_{ij} = y_{\#}$ | Shuffle labels for experimental conditions, so that it is unclear which set of results matches which conditions. | Testing a prediction that hard-copy books are better comprehended than audiobooks. |
| Item scrambling $\theta_{ij} = y_{\#\#}$ | Randomly relabel each data point to de-identify experimental conditions. | Analysing group differences that might be easy to recognize even with noise and bias (for example, effects of neighbourhood and school on crime victimization). |
| Various combinations | Row scrambling: keep pairs of variables together to preserve correlation. Variable blinding: swap labels of various variables. | |

y_{ij} is the i th observation in the j th condition ('cell') of the study; β_k is the k th parameter of a model; θ_{ij} is y_{ij} or β_k after blinding; n_{ij} is random error, b_j is a bias term, and $\#$ denotes a randomly swapped subscript.

strategy (see 'Blinding strategies').

A fertile approach is to present panels of possible results, in which the real results may or may not be interspersed among various decoys. Such a blinded presentation of possibilities typically triggers useful questions. For example, a plausible, although still blinded, graph may lead the researcher to ask whether a sample explores the full range of an independent variable, or it might trigger a revisiting, before unblinding, of the scaling of one of the variables. Another graph might suggest that the whole effect is driven by a single outlier point, and suggest that the researcher needs more data, again before unblinding. Often, a panel can seem implausible until the investigator recognizes an assumption that, if wrong, would produce such a pattern.

COMMON OBJECTIONS

Blind analysis is not a panacea, but it is much more feasible than many think. Here we address common objections.

Won't people just peek at the raw data?

Blind analysis is not immune to fraud. But in ordinary research, teams of investigators can help to enforce compliance. Where blinding is part of the culture, graduate students and postdocs often become its most effective guardians, for example, flagging the risk if their adviser asks for a plot that might accidentally unblind the result.

Can't we avoid bias another way? Other solutions have been proposed, including pre-registered analysis plans, cross-lab replication, the p -curve, adversarial collaboration, Bayesian analytical methods and sensitivity analysis⁶. These techniques all have their place, but they do not fully address the specific problem. For example, preregistration requires that data-crunching plans are determined before analysis, and offers some of the same benefits as blind analysis. But

it also limits the scope of analysis. Because many analytical decisions (and computer programming bugs) cannot be anticipated, investigators will be forced to make some decisions knowing (consciously or unconsciously) how their choices affect the results. Blind analysis enables the investigator to engage in analysis, exploration and finalization without worrying about such bias.

Isn't blind analysis too much hassle? There is extra effort involved. Often the analyses that at first seem most worth the trouble are those that involve expensive data, high-stakes decisions or topics especially prone to bias. However, blinding analyses could be as simple as asking a colleague down the hall to scramble labels. And when safety is at stake, such as in some clinical trials, it often makes sense to set up an unblinded safety monitor while the rest of the analytical team is in the dark⁷. Technology could help here: an important advance would be the introduction of off-the-shelf algorithms in standard analysis software to maintain the blinding until the group is ready to reveal the results. A less obvious benefit is the sheer fun of the dramatic moment when the results are revealed.

Won't blinding lose outcomes that depend on analyses done once the result is seen?

Ideally, among the panels of blinded results there is also the set of actual results, so the researchers could use this to consider further implications (along with the implications of the other hypothetical results). Of course, there will still be post-hoc (post-unblinding) discussion; it will simply be possible to distinguish work investigators performed while still unaware of the results.

MAKING IT HAPPEN

We see two challenges for the widespread dissemination of blind analysis. The first is technical: learning to blind what should be

blinded while preserving features needed to permit appropriate analysis. The second is motivational: creating incentives for investigators to adopt a method that might make it harder for them to come up with desirable (although possibly false) results.

Supplementary research grants that encourage testing blind-analysis methods across multiple fields could help to tackle both challenges. The efficacy of various approaches — methods of blinding, pre-registration and other measures against confirmation bias — should be treated as empirical questions to be answered by future research, as demonstrated by a 2015 study of the effects of preregistration⁸. Many blinding techniques have already been developed², and hopefully, a meta-science of best practices will emerge.

Wider use of blinded analysis could be a boon to the scientific community. The main use is to filter out biased inferences, but there are other benefits, too. First, blind analysis can help investigators to consider the opposite of their expectations, a proven strategy for sound reasoning⁹. Second, blinding exposes the investigator to unexpected patterns that fuel both creativity and scrutiny of the theory and methodology¹⁰.

Finally, blind analysis helps to socialize students into what sociologist Robert Merton called science's culture of 'organized scepticism'. As Feynman put it: "This long history of learning how to not fool ourselves — of having utter scientific integrity — is, I'm sorry to say, something that we haven't specifically included in any particular course that I know of. We just hope you've caught on by osmosis. The first principle [of science] is that you must not fool yourself — and you are the easiest person to fool." ■

Robert MacCoun is a psychologist and a professor of law at Stanford University in California, USA. **Saul Perlmutter** is a professor of physics at the University of California, Berkeley, USA. He shared the 2011 Nobel Prize in Physics.
e-mails: rmaccoun@stanford.edu; saul@lbl.gov

1. Feynman, R. P. *Surely You're Joking, Mr. Feynman!* (W. W. Norton, 1985).
2. Klein, R. J. & Roodman, A. *Annu. Rev. Nucl. Part. Sci.* **55**, 141–163 (2005).
3. Conley, A. et al. *Astrophys. J.* **644**, 1–20 (2006).
4. Miller, L. E. & Stewart, M. E. *Contem. Clin. Trials* **32**, 240–243 (2011).
5. MacCoun, R. J. *Annu. Rev. Psychol.* **49**, 259–287 (1998).
6. Miguel, E. et al. *Science* **343**, 30–31 (2014).
7. Meinert, C. L. N. *Engl. J. Med.* **338**, 1381–1382 (1998).
8. Kaplan, R. M. & Irvin, V. L. *PLoS ONE* **10**, e0132382 (2015).
9. Lord, C. G., Lepper, M. R. & Preston, E. J. *Pers. Soc. Psychol.* **47**, 1231–1243 (1984).
10. Simonton, D. K. *Rev. Gen. Psychol.* **15**, 158–174 (2012).



Many hands make tight work

Crowdsourcing research can balance discussions, validate findings and better inform policy, say
Raphael Silberzahn and Eric L. Uhlmann.

Our experience with crowdsourced analysis began in 2013, shortly after we published research¹ suggesting that noble-sounding German surnames, such as König (king) and Fürst (prince), could boost careers. Another psychologist,

Uri Simonsohn at the University of Pennsylvania in Philadelphia, asked for our data set. He was sceptical that the meaning of a person's name could affect life outcomes. While our results were featured in newspapers around the world, we ▶

► awaited Simonsohn's response.

Re-running our analysis yielded the same outcome. But Simonsohn's different (and better) analytical approach showed no connection between a surname such as Kaiser (emperor) and a job in management. Despite our public statements in the media weeks earlier, we had to acknowledge that Simonsohn's technique showing no effect was more accurate. To make this finding public, we wrote a commentary with Simonsohn, in which we contrasted our analytical approaches and presented our joint conclusion².

In analyses run by a single team, researchers take on multiple roles: as inventors who create ideas and hypotheses; as optimistic analysts who scrutinize the data in search of confirmation; and as devil's advocates who try different approaches to reveal flaws in the findings. The very team that invested time and effort in confirmation should subsequently try to make their hard-sought discovery disappear.

We propose an alternative set-up, in which the part of the devil's advocate is played by other research teams.

THE EXPERIMENT

Last year, we recruited 29 teams of researchers and asked them to answer the same research question with the same data set. Teams approached the data with a wide array of analytical techniques, and obtained highly varied results. Next, we organized rounds of peer feedback, technique refinement and joint discussion to see whether the initial variety could be channelled into a joint conclusion. We found that the overall group consensus was much more tentative than would be expected from a single-team analysis³.

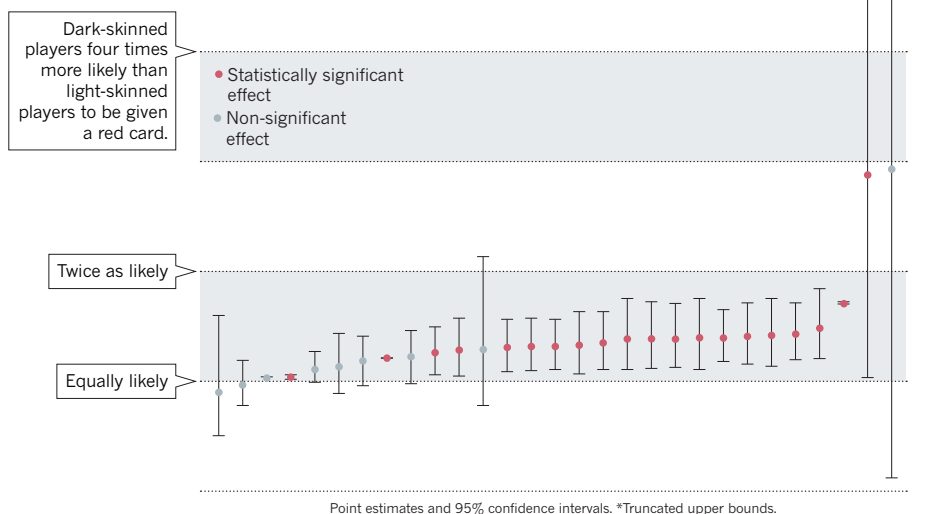
The experience convinced us that bringing together many teams of skilled researchers can balance discussions, validate scientific findings and better inform policymakers. Here, we describe how such a crowdsourcing approach can be a useful addition to research.

In many academic disciplines, multiple teams work with the same data set, for instance the World Values Survey data in political science or genome databases in genetics research. However, each team is typically keen to investigate its own questions and search for new phenomena. Thus hypotheses and results are often held close. Only after a conclusion is ready for presentation are methods and outcomes shared with other researchers, leaving limited opportunity for critical discussion.

By contrast, our project set out to enable researchers to exchange methods and refine analyses before forming their conclusions. We asked the teams to approach the same data with the same question: are

ONE DATA SET, MANY ANALYSTS

Twenty-nine research teams reached a wide variety of conclusions using different methods on the same data set to answer the same question (about football players' skin colour and red cards).



football (soccer) referees more likely to give red cards to players with dark skin than to players with light skin? This question touches on broad issues, such as how prejudice affects sports and how well the effects of prejudice, as detected in laboratory settings, show up in the real world.

Together with psychologist Brian Nosek, director of the Center for Open Science in Charlottesville, Virginia, and Dan Martin, a graduate student in quantitative psychology at the University of Virginia in Charlottesville, we developed a crowdsourcing methodology to coordinate analysts' efforts. Researchers who signed up for the project held varied opinions about whether an effect existed.

All teams were given the same large data set collected by a sports-statistics firm across four major football leagues. It included referee calls, counts of how often referees encountered each player, and player demographics including team position, height and weight. It also included a rating of players' skin colour. As in most such studies, this ranking was performed manually: two independent coders sorted photographs of players into five categories ranging from 'very light' to 'very dark' skin tone.

The teams independently tested their hypotheses. Each made its own decisions about how to best analyse the data set. We then took an inventory of all the approaches. Each team provided details such as which statistical model they used — everything from Bayesian clustering to logistic

regression and linear modelling — what variables they used and why. Teams' approaches were then anonymized and sent back to all of the researchers without revealing results.

The researchers were asked to rate the validity of each approach and to provide in-depth feedback on three approaches. Then we sent participants a document listing all the approaches and associated feedback, and gave teams time to update their analyses. Then the groups documented their analyses and models, which, along with the results, were shared with all teams.

After that, we invited all the researchers to discuss the results through e-mail exchanges. Some approaches were deemed less defensible than others, but no consensus emerged on a single, best approach. After the discussion, we gave researchers the chance to add a note to their individual reports in light of others' work (in other words, to express doubts or confidence about their approach). Finally, we presented the teams' findings in a draft manuscript, which the participants were invited to comment on and modify.

DIVERSITY OF RESULTS

Of the 29 teams, 20 found a statistically significant correlation between skin colour and red cards (see 'One data set, many analysts'). The median result was that dark-skinned players were 1.3 times more likely than light-skinned players to receive red cards. But findings varied enormously, from a slight (and non-significant) tendency for referees to give more red cards to light-skinned players to a strong trend of giving more red cards to dark-skinned players. After reviewing each other's reports, most team leaders concluded that a correlation



NATURE.COM
For Nature's special collection on reproducibility, see: go.nature.com/huhbyr

between a player having darker skin and the tendency to be given a red card was present in the data.

Nonetheless, the fact that so many analytical approaches can be presented — and justified — gives researchers and the public a more nuanced view. Any single team's results are strongly influenced by subjective choices during the analysis phase. Had any one of these 29 analyses come out as a single peer-reviewed publication, the conclusion could have ranged from no race bias in referee decisions to a huge bias.

Most researchers would find this broad range of effect sizes disturbing. It means that taking any single analysis too seriously could be a mistake, yet this is encouraged by our current system of scientific publishing and media coverage.

PROS AND CONS

For many research problems, crowdsourcing analyses will not be the optimal solution. It demands a huge amount of resources for just one research question. Some questions will not benefit from a crowd of analysts: researchers' approaches will be much more similar for simple data sets and research designs than for large and complex ones. Importantly, crowdsourcing does not eliminate all bias. Decisions must still be made about what hypotheses to test, from where to get suitable data, and importantly, which variables can or cannot

be collected. (For instance, we did not consider whether a particular player's skin tone was lighter or darker than that of most of the other players on his team.) Finally, researchers may continue to disagree about findings, which makes it challenging to present a manuscript with a clear conclusion. It can also be puzzling: the investment of more resources can lead to less-clear outcomes.

Still, the effort can be well worth it. Crowdsourcing research can reveal how conclusions are contingent on analytical choices. Furthermore, the crowdsourcing framework also provides researchers with a safe space in which they can vet analytical approaches, explore doubts and get a second, third or fourth opinion. Discussions about analytical approaches happen before committing to a particular strategy. In our project, the teams were essentially peer reviewing each other's work before even settling on their own analyses. And we found that researchers did change their minds through the course of analysis.

Crowdsourcing also reduces the incentive for flashy results. A single-team project may be published only if it finds significant effects; participants in crowdsourced projects can contribute even with null findings. A range of scientific possibilities are revealed, the results are more credible and analytical choices that seem to sway conclusions can point research in fruitful directions. What is more, analysts

learn from each other, and the creativity required to construct analytical methodologies can be better appreciated by the research community and the public.

Of course, researchers who painstakingly collect a data set may not want to share it with others. But greater certainty comes from having an independent check. A coordinated effort boosts incentives for multiple analyses and perspectives in a way that simply making data available post-publication does not.

The transparency resulting from a crowdsourced approach should be particularly beneficial when important policy issues are at stake.

“Under the current system, strong storylines win out over messy results.”

The uncertainty of scientific conclusions about, for example, the effects of the minimum wage on unemployment, and the consequences of economic austerity

policies should be investigated by crowds of researchers rather than left to single teams of analysts.

Under the current system, strong storylines win out over messy results. Worse, once a finding has been published in a journal, it becomes difficult to challenge. Ideas become entrenched too quickly, and uprooting them is more disruptive than it ought to be. The crowdsourcing approach gives space to dissenting opinions.

Scientists around the world are hungry for more-reliable ways to discover knowledge and eager to forge new kinds of collaborations to do so. Our first project had a budget of zero, and we attracted scores of fellow scientists with two tweets and a Facebook post.

Researchers who are interested in starting or participating in collaborative crowdsourcing projects can access resources available online. We have publicly shared all our materials and survey templates, and the Center for Open Science has just launched ManyLab, a web space where researchers can join crowdsourced projects. ■

Raphael Silberzahn is assistant professor in the Department of Managing People in Organizations at IESE Business School, Barcelona, Spain. **Eric L. Uhlmann** is associate professor of organizational behaviour at INSEAD in Singapore. e-mails: rsilberzahn@iese.edu; eric.uhlmann@insead.edu

1. Silberzahn, R. & Uhlmann, E. L. *Psychol. Sci.* **24**, 2437–2444 (2013).
2. Silberzahn, R., Simonsohn, U. & Uhlmann, E. L. *Psychol. Sci.* **25**, 1504–1505 (2014).
3. Silberzahn, R. et al. Preprint available at <https://osf.io/j5v8f> (2015).



Mario Balotelli, playing for Manchester City, is shown a red card during a match against Arsenal.

MICHAEL REGAN/GETTY



Alfred Wegener crossing a glacier on his final, fatal expedition to Greenland, in 1930.

GEOLOGY

The continental conundrum

Ted Nield hails a biography of Alfred Wegener, who proposed the theory preceding plate tectonics.

There is a moment in *Star Wars: The Empire Strikes Back* (1980) that neatly encapsulates two contrasting scientific types. An officer reports a lead from a drone sent to find a rebel hideout. Admiral Ozzel dismisses his claim with statistical objections; Darth Vader takes one look and says, "That's it. The rebels are there." Such intuitive certainty characterized the discoverer of continental drift, Alfred Wegener, subject of science historian Mott Greene's much-anticipated biography. Wegener saw what others missed and knew that he was right — and mostly was. The force was with him.

Every geoscientist has heard of Wegener (1880–1930). Yet we have waited 85 years for a biography to explain who he was and what he achieved beyond the one thing that made him immortal. In this vacuum, myths have proliferated. Almost everything most geologists think they know about Wegener (that as a

'meteorologist' he was an 'unknown outsider', for instance) is wrong. Greene beautifully puts the record straight with a portrait of Wegener as a respected 'cosmic physicist': trained in astronomy, Wegener applied physics to the observable Universe, from stars and planets to Earth's atmosphere, crust and interior.

Wegener was born in Berlin, to Anna Schwarz and clergyman, classicist and pedagogue Richard Wegener. His childhood was one of hard work and outdoorsy self-discipline. Sharing his lifelong need for physical exertion with his older brother and fellow explorer Kurt, Wegener developed a love of ballooning that led to their 1906 world record of 52.5 hours of continuous flight. His passion for the outdoors led him towards understanding the natural world through physics.

The tenor of his life was set. Wegener made three gruelling expeditions to Greenland — which Greene graphically recounts. His dogged pursuit of scientific data, notably on the structure of the upper atmosphere and the zones of shear between its component layers, verged on the superhuman.

Wegener's way of working became evident as early as 1909.

In a paper about the layered structure of the atmosphere, he took others' published data and made connections they had missed to propose a simpler, more elegant conclusion. Similarly, in 1911, Wegener made his wild surmise about the westward drift of the Americas, after first looking into Richard Andree's great work, his world map *Allgemeiner Handatlas* (1881). Wegener was not the first to notice the fit between the Atlantic Ocean's opposing shores — more specifically, between the continental-shelf margins of South America and Africa. But in 1912, he was the first to hypothesize that the Atlantic is young, formed as the Americas drifted away from Europe and Africa — an idea developed more fully in 1915 (see also M. Romano and R. L. Cifelli *Nature* 526, 43; 2015). He also posited that all continents had once been grouped as a supercontinent, Pangaea.

The contemporary orthodoxy proposed by Austrian geologist Eduard Suess held that the Earth was shrinking, causing parts of continents to founder and marooning populations of similar animals in widely separated lands. Wegener recognized this as physically impossible. The principle of isostasy — developed by nineteenth-century geophysical theorists John Henry Pratt, George Biddell Airy and Osmond Fisher — had shown that rafts of continental rock 'float' in the denser material of the Earth's mantle; continents cannot 'sink'. To Wegener, it was more elegant to assume that the continents moved laterally. The fit of the Atlantic coastlines was, for him, too convincing to be a coincidence. How they moved was something for future research.

This was the thinking of a physicist. Yet initially, most physicists (and geologists, including influential figures such as US Geological Survey engineer Bailey Willis) rejected the idea. Geologists in Europe and South Africa converted first, perhaps because their wider knowledge of global geology led them to stronger supporting evidence. Most US scientists held out until the 1950s, usually citing a lack of mechanism. But their scepticism had more to do with the culture of US science, which, in contrast to Wegener's approach,



Alfred Wegener: Science, Exploration, and the Theory of Continental Drift

MOTT T. GREENE
Johns Hopkins University Press: 2015.

ULLSTEIN BILD VIA GETTY

➔ **NATURE.COM**
For more on science in culture see:
nature.com/booksandarts

relied on considering multiple hypotheses. The vehemence of the debate hints, too, that their rejection was more of Old World hierarchies than of Wegener's theory.

Indeed, the barriers were largely sociological, as Greene shows. Wegener was born at the wrong time. The First World War interrupted his career, and began the long isolation of German scientists just when his great idea most needed discussion. He also fell between disciplinary stools. He published important contributions on astronomy, meteoritics, atmospheric science, climatology, palaeoclimatology, geology, geophysics, geodesy and glaciology. When he died aged just 50 in 1930, from a heart attack on the Greenland ice sheet, he left no disciples. 'Cosmic physics' broke up like a supercontinent.

As a result, Wegener makes a challenging subject, which Greene tackles through extensive archival research, travel and circumstantial evidence. Wegener left no extensive notebooks; much of his unpublished writing was destroyed by war or neglect, and he was not given to personal revelation. Like many polar explorers, he was wrapped up in his work. It seems amazing that Wegener married, until we realize that his wife, Else, was the daughter of his collaborator, émigré Russian climatologist Wladimir Köppen.

Others have covered aspects of the enigmatic geoscientist's legacy. Henry Frankel published the comprehensive four-volume *The Continental Drift Controversy* in 2012 (Cambridge Univ. Press); Naomi Oreskes expertly explored US opposition to the theory in *The Rejection of Continental Drift* (Oxford Univ. Press, 1999). Greene's full picture of the man is set masterfully within the wider development of the subjects on which he exerted influence (or failed to). If this wonderful book has a weakness, it is a dearth of illustrations; but those on show include many previously unpublished expedition pictures.

Following the advice of Michael Faraday's biographer L. Pearce Williams, Greene has "read everything his subject wrote, everything he read, and as much as possible of what the people he read, read". He has also travelled everywhere Wegener went, including Greenland. The labour has taken more than 20 years. The result is a magnificent, definitive and indefatigable tribute to an indefatigable man. ■

Ted Nield is the author of *Supercontinent*. His latest book is *Underlands*. e-mail: ted.nield@geolsoc.org.uk

SCIENCE FICTION

Crusoe on Mars

Elizabeth Gibney relishes Ridley Scott's disco-laced chronicle of survival on the Red Planet.

Watching *The Martian* might disturb your dreams, but it will not give you nightmares. Veteran director Ridley Scott's Mars is dirty, rugged and perilous, yet the lingering impression is of a planet more rose-tinted than red.

The Martian tells the story of NASA astronaut and botanist Mark Watney (Matt Damon), stranded alone on Mars after his crew is forced into an emergency evacuation. Believing him dead, the team heads back to Earth, leaving Watney to work out how to survive until NASA can launch a rescue mission. Despite being faced with unimaginable loneliness, and probable death, Damon's Watney seems bizarrely chipper. This danger-riddled film is almost absurdly fun.

Cheeriness in the face of peril is a new turn for Scott. Bar his much-derided *A Good Year* (2006), the director of *Gladiator* (2000) and *Alien* (1979) does not really do comedy. *The Martian* inherits much of its wit from the source novel, originally self-published by writer Andy Weir in 2011. Its irreverence also owes much to scriptwriter Drew Goddard, whose work includes television's *Buffy The Vampire Slayer*. Thanks in large part to Damon's likeably wry portrayal of Watney, the tone works. You are happy to share his company for the duration, particularly given the joyful, mostly disco soundtrack.

In his crew's habitat, or 'Hab', Watney has an array of solar panels and apparently unending rolls of duct tape. But food is limited, and to grow his own he must source water (he manages by burning hydrazine from rocket fuel). After last year's Christopher Nolan blockbuster *Interstellar*

(Z. Merali *Nature* 515, 196–197; 2014), Damon might seem to be repeating the role of abandoned astronaut, but there are few parallels. *The Martian* is strong on the workings of science — logic, problem-solving and perseverance — in contrast to Nolan's rather pompous affair. Ridley gives us failures as well as the triumphs — and

includes geeky shout-outs to the second law of thermodynamics, the hexadecimal

system and radioisotope thermoelectric generators. There is also Watney's heavily trailed quote that the only way to survive is to "science the shit out of this".

Some press reports have focused on Weir and Scott's efforts to get the science right. And it is true that if humans were to go to Mars — a NASA goal for the 2030s — the set-up would probably resemble that in the film. Missions are run in stages, with equipment sent ahead. The crew dawdles on space trips many months long, and once on Mars, scrabbles to collect the perfect soil sample. Yet the science is hardly faultless. Mars's low gravity would be more visible in the astronauts' gait. The thin atmosphere would make radiation one of Watney's main concerns. But when human spirit and ingenuity are the heart of the film, dwelling on these elements seems unfair.

Given *The Martian*'s fantastic cast, it is a shame that only Damon is allowed to shine. Kristen Wiig, as NASA's director of media relations, too often stands wide-eyed in the background when the public-relations roller coaster of the plot should have brought her character to life. Jeff Daniels as the NASA administrator could have been a cynical counterbalance to the pure-as-snow researchers, but he, too, is one of the good guys.

The film has been praised for its diversity, and there appear to be more women and ethnic-minority staff at *The Martian*'s fictional NASA than in the real agency. Chiwetel Ejiofor enjoys himself playing Vincent Kapoor, head of Mars missions. Jessica Chastain is perfect as serious but warm-hearted mission-commander Melissa Lewis. There are two female crew members, both worthy role models; but whereas several of the male astronauts have children, neither of the women does — perhaps a disappointing hint that space and motherhood do not go together.

The Martian is more feel-good than 2013's *Gravity*, and lighter than *Interstellar*. Watney's psyche is left largely unexplored. Where the other two blockbusters invite the audience to stand in awe of nature as a mighty beast, *The Martian* asks you to saddle it. Enjoy the ride. ■

The Martian
RIDLEY SCOTT
20th Century Fox:
2015.

Matt Damon in
The Martian.

Elizabeth Gibney reports on physical sciences for *Nature* from London.





Delegates at the 2011 UN Climate Change Conference in Durban, South Africa.

ENVIRONMENT

Climate stalemate

Oliver Geden welcomes an analysis of the political inertia impeding a global treaty to limit warming.

This year's climate summit in Paris, starting on 30 November, is a focal point for policymakers, diplomats, political analysts and scientific advisers. Each week of the run-up brings sweeping political declarations and comprehensive reports. Recommendations abound for a global climate treaty to avoid dangerous climate change.

The United Nations Framework Convention on Climate Change (UNFCCC) passed more than 20 years ago, and the Intergovernmental Panel on Climate Change (IPCC) has published 5 assessment reports. Yet there is still no effective regime for emissions reductions and adaptation to climate change.

It is time to consider the workings of international politics to help to explain why the UN has not been able to deliver a comprehensive, ambitious climate treaty. That would rein in expectations for the 21st Conference of the Parties (COP21) in Paris. Such analysis is rare, making *Climate Change in World Politics*, by international-relations scholar John Vogler, most welcome. Vogler provides a detailed reconstruction of how the international community has dealt with climate change over the past 25 years, and how that has contributed to the current impasse.

The debate on global climate governance has focused more on declarations of intent from key state and non-state actors. Less attention has been given to results delivered by the network of international institutions, centred on the UNFCCC and its various

agreements. In analyses of climate policy, the dominant, 'functionalist' approach looks at forums of international cooperation such as the UNFCCC as if they exist solely to serve collective efforts towards climate solutions. Vogler begs to differ.

He proceeds from two premises. First, sovereign states are still the main determiners of global climate policy, because only they have the necessary regulatory authority. Second, to assess the potential and limits of such policy accurately, it is necessary to consider power politics, prestige-seeking and other behaviours specific to national governments. Vogler rightly assumes that most activities under global climate policy do "not necessarily accord" with the stated purpose of mitigation and adaptation. Regional conflicts or the formation of factions, such as the Group of 77 (G-77) caucus of developing countries, can lead to long-lasting coalitions. In the G-77, development policy generally trumps climate policy. And for a long time, emerging economies such as Brazil and China continued to present themselves as developing countries, effectively avoiding mitigation commitments. State governments are also driven by

"Some key drivers of emissions have been continually exempted from the scope of international climate policy."

Climate Change in World Politics
JOHN VOGLER
Palgrave Macmillan:
2015.

short-term economic interests, aimed at maintaining domestic political power.

So although governments like to tout their problem-solving capabilities, their approach is often more vague: one of 'dealing with problems'. Vogler shows that there were many competing definitions of climate change even before the UNFCCC was passed. And he reveals how some of the principles that define the climate regime gained widespread acceptance — for instance, the focus on territorial emissions and the principle that industrialized countries have the greatest responsibility for mitigation. He also shows how some key drivers of emissions (population growth and globalized trade flows in particular) have been continually exempted from the scope of international climate policy.

At the same time, the fragmented portfolio of themes covered by the climate regime (for example, environmental justice, carbon markets and climate refugees) has emerged from struggles for power. Governments wage these battles through complicated procedural rules and selective perceptions of climate research. This can be seen, for instance, in the variable interpretations of cumulative historical emissions of industrialized countries and emerging economies.

Vogler pinpoints other reasons for pessimism about COP21. Many countries — notably the United States and China — have no interest in an effective top-down regime or the strong international institutions needed to prevent dangerous climate change. No deal will be reached in Paris that will make it possible to limit the increase in global temperatures to 2°C above preindustrial levels, even though this was the goal of the Durban Platform for Enhanced Action, decided at COP17 in 2011.

The emerging climate regime, with nationally determined mitigation at its core, does mark a shift towards respect for the global power structure. Nevertheless, COP21 is likely to confirm the modus operandi of UN climate policy. That is: "kicking the can down the road in order to delay potentially difficult and costly decisions".

Vogler's intention is not to indict the key actors in climate policy, nor to accuse them of a lack of political will. His aim is to warn against unrealistic expectations of the problem-solving capacity of such a complex and fragmented form of policy coordination. Already, such expectations have led to disillusionment and threaten a rejection of the entire process. ■

Oliver Geden is head of the European Union Research Division at the German Institute for International and Security Affairs (SWP) in Berlin.
e-mail: oliver.geden@swp-berlin.org

Correspondence

Pricing would limit carbon rebound

Unilateral national climate policies are not strict enough to control carbon rebound — a side effect of some energy-conservation strategies that undercuts net carbon savings. I suggest that a global agreement on variable carbon pricing at the forthcoming climate summit in Paris would reap considerable rebound-related benefits.

Economy-wide studies indicate that overall carbon rebound is at least 50%, depending on the country (J. Dimitropoulos *Energy Policy* **35**, 6354–6363; 2007). Despite this, the effects of rebound have been largely ignored by the Intergovernmental Panel on Climate Change and at United Nations climate meetings.

Technical standards do not control rebound effectively: they cover only a small subset of products. For example, when the European Union began phasing out incandescent light bulbs in 2009, light-emitting diodes became so widespread that any energy savings were reduced.

The most effective way to discourage rebound is through carbon pricing, a policy that underpins all potential energy-savings decisions. Any rebound tendency would elicit a higher carbon price under a cap-and-trade permit scheme. A carbon tax would require frequent adjustment to achieve the same outcome. This would be difficult politically, especially in the form of nationally distinct taxes.

Jeroen C. J. M. van den Bergh
ICREA, Barcelona; Autonomous University of Barcelona, Spain; and VU University Amsterdam, the Netherlands.
jeroen.bergh@uab.cat

Safeguard the ideas of junior scientists

Good institutional practice fosters research reproducibility (see C. G. Begley *et al. Nature* **525**, 25–27; 2015). But individuals

still have a responsibility to work with their institutes by practising research with integrity.

Scientists are encouraged to be open about their research, which makes unpublished junior scientists particularly vulnerable to intellectual-property theft. This can be a risk when seeking out collaborations, presenting at conferences and submitting manuscripts for review, and even after employment interviews.

We are all subject to the same frailties and pressures, but we are members of a community with a common goal. We cannot always own originality, but we must show respect for it.

Jennifer S. Le Blond *Imperial College London, UK.*
j.le-blond@imperial.ac.uk

Diesel pollution long under-reported

The furore over Volkswagen's cheating of US emissions tests (see *Nature* <http://doi.org/723>; 2015) prompts a reminder that pollution from diesel vehicles has long been under-reported. This includes nitrogen oxides, hydrocarbons and particulates.

Emissions can be measured across cities using instruments on aircraft and high towers, and for vehicles using number-plate recognition and remote sensing. Modern diesel engines emit roughly four times more nitrogen oxides on average than are recorded in lab tests, which use unrepresentative driving cycles and technical strategies to reduce emissions (D. C. Carslaw and G. Rhys-Tyler *Atmos. Envir.* **81**, 339–347; 2013). Real-world emissions of diesel hydrocarbons exceed estimates used for air-quality planning by up to 70 times (R. E. Dunmore *et al. Atmos. Chem. Phys.* **15**, 9983–9996; 2015).

Particulate matter from diesel is estimated to kill 29,000 people each year in the United Kingdom (see go.nature.com/kdzn4r). This figure will rise when the UK Committee on the Medical

Effects of Air Pollutants, which advises government, quantifies the extra health burden associated with nitrogen dioxide.

Improvements in urban air quality stalled a decade ago in many European cities, where nitrogen dioxide often exceeds regulatory standards and global health guidelines. To tighten up diesel-emissions control, tests need to be more accurate, more transparent and regulated more rigorously (see also F. J. Kelly and J. C. Fussell *Envir. Geochem. Health* **37**, 631–649; 2015).

Alastair C. Lewis, David C. Carslaw *University of York, UK.*
Frank J. Kelly *King's College London, UK.*
ally.lewis@york.ac.uk

Tackling soil loss across Europe

A European Commission analysis indicates that soil erosion continues to outstrip soil formation across the European Union, but that the Common Agricultural Policy is narrowing the gap (P. Panagos *et al. Environ. Sci. Policy* **54**, 438–447; 2015).

The amount of soil lost to water erosion in Europe equates to an estimated economic loss of about US\$20 billion per year, based on a replacement cost of \$20 per tonne. Between 2000 and 2010, intervention measures through the Common Agricultural Policy have reduced the rate of soil loss in the European Union by an average of 9.5% overall, and by 20% for arable lands.

Continued monitoring of human-induced changes to soil every 5–10 years will be crucial for refining soil policies (D. A. Robinson *Science* **347**, 140; 2015).

Panos Panagos, Pasquale Borrelli *European Commission Joint Research Centre — IES, Ispra, Italy.*

David A. Robinson *NERC Centre for Ecology and Hydrology, Environment Centre Wales, Bangor, UK.*
panos.panagos@jrc.ec.europa.eu

Protect biodiversity, not just area

The Convention on Biological Diversity's Aichi Target 11 mandates that 17% of terrestrial and 10% of marine environments be conserved in protected areas by 2020. Such simple numeric indicators act as motivators and a measure of progress. But striving to meet the stipulated coverage should not compromise the convention's broader goal of maximizing biodiversity.

Area coverage is the only element of Target 11 that is on track, at least on land (D. P. Tittensor *et al. Science* **346**, 241–244; 2014). Other crucial elements are effective, equitable biodiversity management; ecological representation of a mix of ecosystems; and connectivity between sites to allow species dispersal. Some species and ecosystems may be lost if implementation of these elements is delayed much longer.

Focusing on area coverage alone risks creating perverse outcomes. It encourages the proliferation of large protected areas that are under little threat, and neglects areas where protection is most needed (see go.nature.com/o5ny9j and go.nature.com/hi6qn5). If not considered in the context of other elements of Target 11, maximizing the area under protection increases the financial and political cost of meeting the same biodiversity goals. As with other global policy goals (see S. Fukuda-Parr *J. Hum. Dev. Capab.* **15**, 118–131; 2014), the abstract global target has created unintended consequences for national conservation planning.

With negotiations beginning in 2016 for the next tranche of the convention's targets, new incentives are needed to emphasize the pivotal additional elements of Target 11.

Megan Barnes* *University of Queensland, St Lucia, Australia.*
megan.barnes@uq.edu.au
*On behalf of 4 correspondents (see go.nature.com/d1vieb for full list).

Eric H. Davidson

(1937–2015)

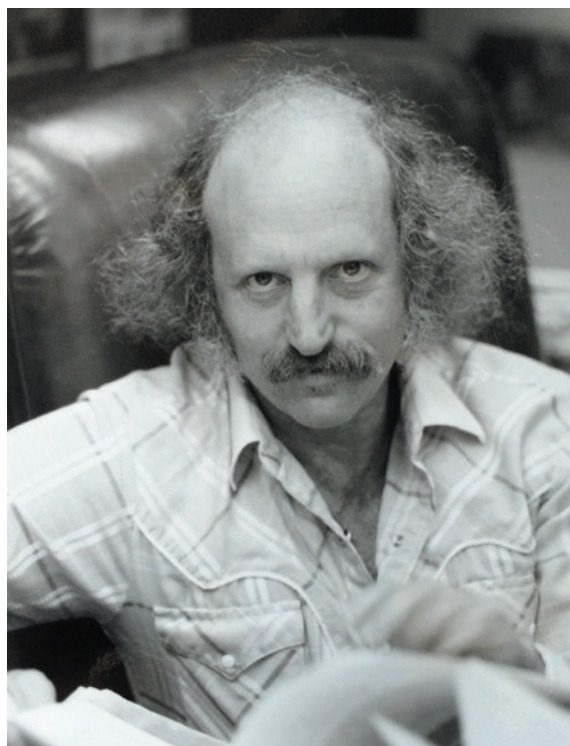
Systems biologist who described gene regulatory networks.

Eric Harris Davidson spearheaded many of the advances that led to our current understanding of how organisms are made from genomes. He helped to show how the coordinated expression of a whole suite of genes determines what progenitor cells specialize into during development. He also helped to pioneer the idea of gene regulatory networks — systems of interacting genes made up of multiple feedback loops, or subcircuits, each performing a specific job in the regulation of gene activity in the cell.

Davidson, who died on 1 September, was born in New York City in 1937. He found his way into research during his teenage years, working with cell physiologist L. V. Heilbrunn at the Marine Biological Laboratory (MBL) in Woods Hole, Massachusetts. During the academic year, Heilbrunn worked on calcium signalling at the University of Pennsylvania in Philadelphia, but he spent the summer at the MBL where he could study cellular processes in marine organisms such as starfish and sea urchins.

It was Davidson's father, Morris, a renowned abstract painter, who had brought Eric and Heilbrunn together. Morris, who ran a summer art school in Provincetown at the tip of Cape Cod in Massachusetts, knew Heilbrunn's wife, Ellen, a painter and teacher. The story goes that Davidson initially joined the MBL to wash dishes but Heilbrunn gruffly told him that anyone working there had to have a research project. Aged just 16, Davidson published an abstract in the *Biological Bulletin* on clotting in sand dollars (*Echinarachnius parma*) — a calcium-dependent cellular process that protects the organisms from infection following injury (E. H. Davidson *Biol. Bull.* **105**, 372; 1953).

In 1954, Davidson began a degree in biology at the University of Pennsylvania, where he worked in Heilbrunn's laboratory. On graduating, he wanted to stay in the lab, but Heilbrunn recommended that he switch to studying gene expression, and Davidson went to work with molecular biologist Alfred Mirsky at Rockefeller University in New York City. He received his PhD in 1963 and remained at Rockefeller until 1971 when he moved to the California Institute of Technology in Pasadena.



By the early 1960s, the 'central dogma' of molecular biology — DNA makes RNA makes protein — had been postulated. The idea that the RNA made from DNA contains specific instructions needed to make proteins had also just been confirmed experimentally. The question that Davidson was interested in was how the coordinated expression of multiple genes brings about cellular differentiation.

To investigate, Davidson applied 'solution hybridization' methods, first to frog oocytes and then to other animal models. Here, fragments of DNA in solution are heated to the point at which the two complementary strands of the double helix separate. Each single strand is then allowed to re-anneal with a partner strand floating in the solution; the more copies there are of any particular sequence fragment, the quicker the strands re-anneal. The approach provided a way to analyse genomic sequences as well as RNA transcripts.

A long-term collaboration with physicist-turned-molecular biologist Roy Britten — who had been working on ways to model hybridization kinetics — led to the publication in 1969 of a paper entitled 'Gene regulation for higher cells: a theory' (R. J. Britten and

E. H. Davidson *Science* **165**, 349–357; 1969). Here, the pair presented the first diagram showing how groups of genes could be expressed in coordination during cellular differentiation.

With the development of recombinant DNA technology in the early 1980s, in which fragments of DNA from multiple sources are spliced together, it became possible to dissect the molecular machinery needed to turn a gene on and off. Through a series of experiments in sea-urchin embryos, Davidson and his colleagues showed that protein-coding genes are controlled by what are usually nearby regulatory modules — DNA sequences that serve as binding sites for transcription factors, the proteins that form complexes to control the transcription of DNA into RNA.

In the mid- to late-1990s, Davidson and his collaborators took advantage of advances in DNA-sequencing technology to examine the genomic sequences near genes. Coupled with data on when genes were being turned on, and in what cells, during the early development of sea urchins, they showed how

the sequences near many genes have a regulatory role. This approach rapidly led to the description of gene regulatory networks. Davidson's work was key in establishing that it is through gene interactions playing out in different ways that cells are assigned their fates and organisms are created. He also helped to show that changes in these networks can result in altered morphologies and traits, the raw material for evolution.

The best scientists, in my view, are the ones who can stand fast in the face of bewildering complexity until they see the patterns emerge. Eric was a good example of such a person. He deeply enjoyed embracing the complexity and trying to develop a cogent view. In the third edition of his book *Gene Activity in Early Development* (Academic, 1986), he credits his father with teaching him about 'the ordering of complex perceptions'. As his father taught him, he taught us. ■

Andrew Cameron is a senior research associate emeritus at the California Institute of Technology (Caltech), Pasadena, California, USA. He joined Eric Davidson's group in Caltech's Division of Biology and Biological Engineering in 1984. e-mail: acameron@caltech.edu

BOB PAZ/CALTECH

MALARIA

Fifteen years of interventions

A comprehensive modelling effort has revealed the relative contributions of different malaria-control measures to the massive reductions in disease prevalence that have occurred in Africa between 2000 and 2015. [SEE ARTICLE P.207](#)

JANET HEMINGWAY

In 1978, when I began working on insect-borne diseases, a child died of malaria every six seconds. Today, although we have made great progress, it is unacceptable that a child still dies every minute from this disease. There are an estimated 600,000 deaths annually, the vast majority of which are in sub-Saharan Africa. To drive down the malaria burden further, and ultimately to try to eradicate the disease, we need to be able to attribute the contributions of different interventions and use this information to optimize our efforts. In this issue, Bhatt *et al.*¹ (page 207) provide the first authoritative, data-driven models to estimate the relative impact that combination drug therapies and mosquito-control strategies have had on clinical cases of malaria in Africa since 2000.

Expert opinion on the relative merits of different malaria interventions varies greatly, and over the past half-century there have been major shifts in emphasis and operational implementation of drug- and insecticide-based approaches. The malaria-eradication efforts led by the World Health Organization in the 1960s used both chloroquine drug treatment and indoor residual spraying with DDT (this involves spraying the inside walls of dwellings with the insecticide). The failure of these efforts, which has been variously attributed to drug and insecticide resistance, under-resourcing and lack of political will to support effective implementation, caused a dramatic shift in approach in Africa. Mosquito control was largely abandoned in favour of improving access to prompt treatment with effective drugs, coupled with intermittent preventive drug treatment of particularly vulnerable populations.

In 2000, the scale of activity and the emphasis on different malaria interventions shifted again in Africa. Initially, there was the uptake of bednets impregnated with long-lasting pyrethroid insecticides, with widespread free or subsidized distribution of these nets to 'at risk' populations. Mosquito control was further supplemented in 2005, supported by the US President's Malaria Initiative, which reintroduced indoor residual spraying in 15 high-burden countries across the continent. Failing first-line drug treatments



LOUISE GUBB/CORBIS

Figure 1 | Net effect. Bhatt *et al.*¹ estimate that 68% of the large reduction in clinical cases of malaria in Africa between 2000 and 2015 is attributable to the use of insecticide-impregnated bednets.

were also replaced with artemisinin-based combination therapies (ACTs) in 79 countries by the end of 2013.

Bhatt *et al.* set out to model the impact of these interventions since 2000, collating the largest global published and unpublished data set ever analysed. They were not able to use the number of deaths from malaria for this modelling, because the quantity and quality of these data in different African settings were inadequate. A more accessible measure was the prevalence rate of malaria parasites in children between the ages of 2 and 10, and so the authors used the number of clinical cases of malaria averted, rather than deaths averted, as the output indicator. The collaborating researchers, who include members of some of the world's best disease-modelling groups, used three independent malaria-transmission models to generate counterfactual geospatial maps — that is, they estimated what the malaria parasite prevalence rates would have been without each intervention.

The authors estimate that 663 million

clinical cases of malaria were averted between 2000 and 2015 (Fig. 1). Most of this improvement (68%) was due to bednets, with 22% resulting from ACTs and 10% from indoor residual spraying. Although these numbers are influenced by the timing, speed and scale on which each intervention was introduced, the massive impact of the mosquito-control interventions will come as a surprise to many who believe that improvements have primarily been driven by the introduction of ACTs².

Over the same 15-year period, malaria infection rates have halved, with three-quarters of this improvement occurring in the past decade. Although this massive improvement in malaria control should be applauded, the study provides a timely warning against complacency. The rate of improvement slowed to 5% per year in 2013, and malaria is an infectious disease that could easily resurge. As we move from the era of the United Nations' Millennium Development Goals, which set the targets for malaria and child-mortality reduction that instigated much of this activity, to the

Sustainable Development Goals, which are more oriented towards improving health systems and health services, we need to maintain and enhance malaria-control activities. The target of universal bednet coverage for at-risk populations is still a distant dream, nets distributed before 2012 now need replacing, and all three interventions are increasingly threatened by the development of mosquito resistance to the insecticides or parasite resistance to the drugs. ACT drug resistance is already well documented in southeast Asia, and would be catastrophic if it spread widely in Africa³. However, the greatest threat may come from the rapid increases in pyrethroid resistance that have occurred in the two main African malaria-vector mosquito species⁴.

Drug resistance is detected rapidly, because health workers and patients can immediately recognize a failing treatment. But it is less

obvious when mosquitoes fail to respond to insecticides, and resistance will already be widespread in a mosquito population before there is an associated increase in the numbers of malaria cases and deaths. An international effort is already in place to try to stem the movement of ACT-resistant malaria parasites from Asia to Africa, and a more proactive approach is now needed for insecticide resistance. The urgency of this is underscored by the improved understanding that Bhatt and colleagues have provided of the crucial role that vector control has had in reducing malaria over the past decade, and, by extrapolation, of the role it will need to have if we are to eliminate the disease.

A healthy portfolio of new antimalarial drugs and insecticides is under development, driven in particular by product-development partnerships coordinated by two non-profit

organizations, the Medicines for Malaria Venture and the Innovative Vector Control Consortium. But substantial development, financial, regulatory and policy hurdles are impeding the roll-out of these agents. If we can overcome these, stay ahead of the 'arms race' of parasite and mosquito resistance, develop an effective vaccine to reduce transmission and optimally deploy these interventions, then no child need die from malaria. ■

Janet Hemingway is at the Liverpool School of Tropical Medicine, Liverpool L3 5QA, UK. e-mail: janet.hemingway@lstm.ac.uk

1. Bhatt, S. *et al. Nature* **526**, 207–211 (2015).
2. Taylor Bright, A. & Winzeler, E. A. *Nature* **498**, 446–447 (2013).
3. White, N. J. *et al. Lancet* **353**, 1965–1967 (1999).
4. Hemingway, J. *et al. Lancet* [http://dx.doi.org/10.1016/S0140-6736\(15\)00417-1](http://dx.doi.org/10.1016/S0140-6736(15)00417-1) (2015).

MOLECULAR BIOLOGY

Mediating transcription and RNA export

The finding that the Mediator protein complex contributes to messenger RNA export from the nucleus in yeast adds to a growing list of roles for the complex in regulating transcriptional processes.

JONATHAN D. RUBIN & DYLAN J. TAATJES

Gene transcription is fundamental to all major physiological processes, and defects in its regulation underlie myriad human diseases. Transcription culminates in the export of messenger RNA transcripts from the nucleus to the cytoplasm, where they are translated into proteins. In a paper in *Cell*, Schneider *et al.*¹ use a combination of structural and cell biology, biochemistry, yeast genetics and transcript analyses to describe how this process is regulated by cooperation between the mRNA export machinery and Mediator — a large, multi-subunit protein complex best known for regulating the activity of the enzyme RNA polymerase II (pol II) during the early stages of transcription².

Many factors converge on nascent mRNA transcripts to facilitate their export from the nucleus. One such factor, the TREX-2 protein complex, regulates export through interactions with other complexes, including pol II and the nuclear pore complex (NPC)³, which acts as a gateway for cellular components to exit the nucleus and enter the cytoplasm. However, the mechanisms by which TREX-2 acts are uncertain.

Schneider *et al.* investigated TREX-2 in the brewer's yeast *Saccharomyces cerevisiae*. In yeast

cells lacking the TREX-2 subunit Sac3, the composition of the Mediator complex changed. Specifically, components of the Mediator 'Cdk8

kinase' module failed to associate with the rest of the complex. The authors demonstrated that TREX-2 physically associates with Mediator, and that this association depends on Sac3 and a Mediator subunit implicated in the activation of transcription, Med31.

A series of experiments then showed a functional interdependence between Mediator and TREX-2. In yeast, genes that are in the process of being transcribed associate with the NPC, presumably to facilitate mRNA export to the cytoplasm⁴. Schneider and colleagues demonstrated that, like Sac3, Med31 is required for gene targeting to the NPC, implying that Mediator is involved in mRNA export (Fig. 1). However, in contrast to cells lacking Sac3, mRNA export seemed normal in cells lacking

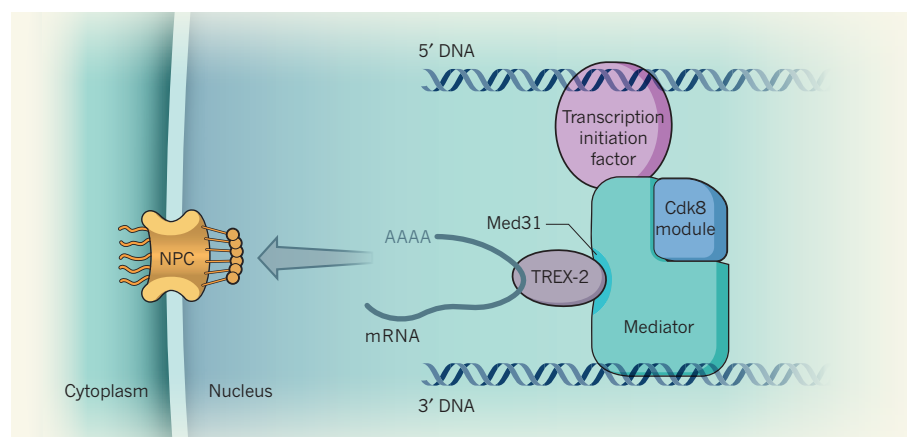


Figure 1 | Complex interactions in messenger RNA export. Following gene transcription, mRNAs are polyadenylated (adenine (A) bases are added to the 3' end of the transcript). Factors involved in the export of mRNA from the nucleus, such as the TREX-2 protein complex, can associate with the mRNA. Schneider *et al.*¹ report that mRNA export is regulated by interactions between TREX-2 and another protein complex, Mediator. In yeast, the Mediator–TREX-2 interaction seems to stabilize Mediator's association with its Cdk8 kinase module — possibly through the TREX-2 subunit Sac3 (not shown) and a subunit of Mediator called Med31. Mediator–TREX-2 interactions help to ensure that genes undergoing transcription are close to the nuclear pore complex (NPC), thereby facilitating mRNA export to the cytoplasm. Gene–NPC interactions are also needed to maintain a function called transcriptional memory, which depends on the formation of a looped gene architecture (only the 5' and 3' ends of a DNA loop are shown) that also seems to be facilitated by Mediator^{7,13}.

Med31 or the Cdk8 kinase protein. Thus, although the authors' data support a role for Mediator in mRNA export, the precise molecular mechanism remains unclear.

Messenger RNA export joins a long list of regulatory roles for Mediator. And Schneider and colleagues' work adds to a growing set of studies^{5–7} suggesting that Mediator regulates late stages of transcription, such as mRNA processing, in addition to its role in transcription initiation².

Note, however, that indirect effects could also contribute to the mRNA export or gene–NPC association defects reported by the authors. Consistent with a previous study⁸, Schneider *et al.* found decreased expression of genes involved in the biosynthesis pathway of sulfur-containing amino-acid residues in cells lacking Sac3 or Med31. An end product of this pathway is S-adenosyl methionine (SAM), which is an essential cofactor for methyltransferase enzymes. A reduction in SAM levels would be expected to inhibit methyltransferase activity, which, in *S. cerevisiae*, regulates mRNA processing and export^{9,10}.

An array of research avenues opens up from Schneider and colleagues' study. For example, the relevance of these yeast findings to human Mediator remains to be determined. Although the authors demonstrated that TREX-2 interacts with Mediator in *S. cerevisiae*, existing studies of human Mediator-interacting proteins are largely devoid of human versions of the TREX-2 subunits. The functional link between *S. cerevisiae* Mediator and mRNA export will probably be evolutionarily conserved in some way, but the mechanisms by which the protein complex acts in human cells are likely to be distinct. Most *S. cerevisiae* mRNAs are not spliced into different versions of the transcript, for instance, in contrast to human mRNAs. Active genes in human cells do not typically associate with the nuclear periphery or the NPC, unlike genes in yeast. Instead, data suggest¹¹ that active genes are found in the nuclear interior in human cells, and that they preferentially associate with structures called PML nuclear bodies, or with 'mobile' NPC proteins such as NUP98 (ref. 4).

An intriguing implication of the current study is the potential involvement of Mediator in regulating transcriptional memory, in which reactivation of specific genes, such as those induced by stress, occurs faster in progeny cells whose parents have previously experienced that stress⁴. This behaviour has been demonstrated in both yeast and mammalian cells⁴, although the underlying mechanisms are incompletely understood. In *S. cerevisiae*, maintenance of gene–NPC associations following cell division correlates with maintenance of transcriptional memory⁴. Schneider *et al.* showed that gene–NPC association requires Med31, suggesting a potential role for Mediator in this process.

An interesting aspect of transcriptional

memory in yeast is its apparent dependence on a looped genomic DNA architecture that juxtaposes the gene's 5' end (the start site for transcription) and 3' end (the site of transcriptional termination)¹². Although DNA architecture at active genes is different in humans, it is noteworthy that Mediator seems to be involved in the formation and stabilization of these loops in both species^{7,13} (Fig. 1). It will be interesting to see whether Mediator, perhaps through its Med31 subunit, contributes to such epigenetic mechanisms of transcriptional memory. For these and other reasons, Schneider and colleagues' study represents an important advance, with mechanistic implications that remain to be explored. ■

Jonathan D. Rubin and Dylan J. Taatjes
are in the Department of Chemistry and

Biochemistry, University of Colorado, Boulder, Colorado 80303, USA.

e-mail: taatjes@colorado.edu

1. Schneider, M. *et al.* *Cell* **162**, 1016–1028 (2015).
2. Allen, B. L. & Taatjes, D. J. *Nature Rev. Mol. Cell Biol.* **16**, 155–166 (2015).
3. Kohler, A., Schneider, M., Cabal, G. G., Nehrbass, U. & Hurt, E. *Nature Cell Biol.* **10**, 707–715 (2008).
4. D'Urso, A. & Brickner, J. H. *Trends Genet.* **30**, 230–236 (2014).
5. Chen, J. *et al.* *RNA* **18**, 2148–2156 (2012).
6. Huang, Y. *et al.* *Mol. Cell* **45**, 459–469 (2012).
7. Mukundan, B. & Ansari, A. J. *Biol. Chem.* **288**, 11384–11394 (2013).
8. Koschubs, T. *et al.* *EMBO J.* **28**, 69–80 (2009).
9. Green, D. M. *et al.* *J. Biol. Chem.* **277**, 7752–7760 (2002).
10. Yu, M. C. *et al.* *Genes Dev.* **18**, 2024–2035 (2004).
11. Therizols, P. *et al.* *Science* **346**, 1238–1242 (2014).
12. Tan-Wong, S. M., Wijayatilake, H. D. & Proudfoot, N. J. *Genes Dev.* **23**, 2610–2624 (2009).
13. Kagey, M. H. *et al.* *Nature* **467**, 430–435 (2010).

NEUROBIOLOGY

Individuality sniffed out in flies

The discovery that certain neurons' odour responses differ between individual fruit flies, but are consistent across the hemispheres of each fly's brain, indicates that sensory processing depends on an individual's experience. [SEE LETTER P.258](#)

THOMAS FRANK & RAINER W. FRIEDRICH

Organisms process sensory information to learn about their environment and to inform future behaviours. The first layers of sensory processing transform information from the sense organs into sparse and stimulus-specific activity patterns across large populations of neurons in the brain¹. But the subsequent processing steps are less well understood. Hige *et al.*² (page 258 of this issue) address this gap in knowledge by studying the olfactory system of the fruit fly *Drosophila melanogaster*.

In insects, olfactory sensory neurons project to a processing centre in the brain called the antennal lobe. At this stage, information about different odours is encoded by overlapping but specific patterns of activity across a relatively small population of projection neurons (about 150 in fruit flies). These activity patterns are transmitted to two higher brain regions, one of which is the mushroom body. Here, odour representations are transformed into sparse activity patterns across a much larger number of neurons called Kenyon cells (flies have around 2,000 Kenyon cells; Fig. 1).

The activity patterns evoked in Kenyon cells by different odours show little overlap. Patterns seem to be generated by random connections to projection neurons, so they vary between

individuals^{3,4}. It is therefore thought that this expansion of odour representations in the mushroom body supports their subsequent classification, for example during learning¹. Indeed, the mushroom body is crucial for learning associations between arbitrary odours and specific values or behaviours. Learning is probably enabled by adjusting the strength of synaptic connections between Kenyon cells and a population of 35 mushroom-body output neurons (MBONs)^{5–7}. The low number of MBONs implies that odour representations are recompressed as they are transferred to higher brain regions, but it remains unclear what information these neurons encode, and how they do this.

Fruit-fly MBONs can be genetically separated into 21 types⁸. Hige *et al.* expressed a fluorescent calcium-sensor protein in most of these cell types, usually targeting one type per fly. Odour-evoked activity of MBONs was measured using a high-resolution multiphoton microscope to detect changes in fluorescence. The authors averaged calcium signals over time and pooled the results across flies to construct mean tuning curves (responses to a set of odours) for each MBON type. They also constructed 'virtual activity patterns' for the whole MBON population by combining the responses of different MBON types.

As observed in other insects^{7,9}, MBONs were

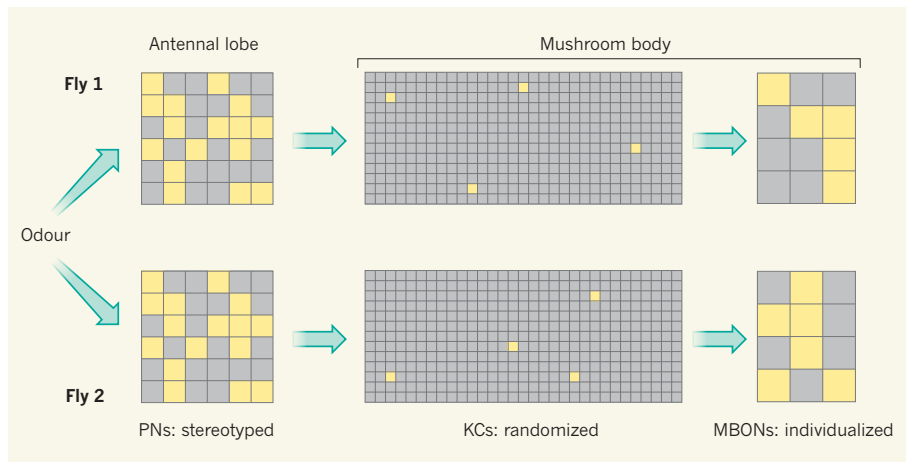


Figure 1 | Reorganizing odour representations. In the antennal lobe of the brains of fruit flies, odours evoke stereotyped activity patterns across roughly 150 projection neurons (PNs; small numbers of each type of neuron are shown, for simplicity). Neurons are represented as squares in a grid, with yellow indicating activity and grey inactivity. The PNs project to a higher brain region called the mushroom body. Here, a population of about 2,000 neurons called Kenyon cells (KCs) integrates inputs from an apparently random combination of PNs, resulting in sparse activity patterns that represent the odour. This large population of KCs then converges onto only 35 mushroom-body output neurons (MBONs). Hige *et al.*² report that the odour responses of MBONs, unlike those of PNs and KCs, are ‘individualized’ — they differ systematically between flies, but are the same in each hemisphere of individual flies (only one hemisphere of each fly is shown).

broadly tuned, each responding to multiple odours. Moreover, the mean tuning curves of different MBON types were often similar. Unlike activity patterns across Kenyon cells, virtual activity patterns across MBONs overlapped and were not well suited for fine odour discrimination. These results suggest that MBONs do not encode odour identity with high efficiency or accuracy. Nevertheless, Hige and colleagues demonstrated that representations of innately attractive odours could be reliably discriminated from repulsive ones. This finding is consistent with the hypothesis¹⁰ that MBONs can transmit information about certain derived properties of the stimulus, such as its valence (positive or negative value).

When interpreting these experiments, two issues should be considered. First, the authors pooled data from several individuals. This approach provides a representative picture when the response of a neuron is stereotyped between individual flies, as in antennal-lobe projection neurons, or when it varies randomly, as in Kenyon cells³. However, pooling is not ideal when responses vary systematically between individuals. Consider the same word written by different people — despite the idiosyncratic handwriting, each word is legible, but the average is a blur. Second, the calcium-imaging method used by Hige *et al.* cannot resolve the fine temporal structure of odour-evoked activity. But in locusts, for example, the precise temporal patterning of MBON activity provides a great deal of information about odour identity⁹.

Hige *et al.* found that the tuning curves of some MBONs were stereotyped across flies, whereas others were more variable. To

explore the source of this variation, they took simultaneous electrical recordings from a Kenyon cell and an $\alpha 2sc$ neuron, the MBON type that showed the greatest variation between individuals. The probability of an excitatory synaptic connection between these cells was only around 30%, suggesting that variable tuning of MBONs is a consequence of variation in the connections they form with Kenyon cells.

Although the tuning curves of $\alpha 2sc$ neurons differed between individuals, the authors showed that the tuning of $\alpha 2sc$ neurons in the two hemispheres of individual flies was nearly identical. Moreover, variability between individuals was abolished in *rutabaga* mutant flies, which have severe learning deficits. The mutation in the *rutabaga* gene disrupts the plasticity process that modifies the strength of synaptic connections. These results suggest that the variation in MBON tuning, unlike that of Kenyon-cell tuning, is not simply caused by randomness in the wiring between Kenyon cells and MBONs, but instead reflects a coordinated process that shapes, or ‘individualizes’, the responses of MBONs in each fly (Fig. 1). Information transfer from Kenyon cells to MBONs might therefore be shaped by experience. Because this process depends on *rutabaga*, the underlying synaptic-plasticity mechanisms might overlap with those involved in associative learning.

Hige and colleagues’ results indicate that MBONs map high-dimensional odour representations in the Kenyon-cell population onto a low-dimensional output. This mapping differs between individuals, perhaps because the precise connectivity between Kenyon cells

and MBONs is shaped by experience. Last year, a behavioural study¹⁰ of fruit flies showed that stimulating different MBONs had specific appetitive or aversive effects that interacted additively, suggesting that the population-wide activity pattern across MBONs represents valence. Hige and colleagues’ results thus indicate that MBONs map representations of odours onto a representation of valence in an experience-dependent fashion. The MBONs then transmit this valence representation to higher brain areas to modulate specific behaviours.

This conclusion is consistent with prevailing views of mushroom-body function^{1,4} and can now be tested further. For example, it may be predicted that spontaneous behavioural reactions to odours are less variable in *rutabaga* flies than in wild-type flies because their odour-to-valence mapping is more stereotyped.

As Marcel Proust highlighted¹¹, the emotions associated with an odour depend on an individual’s experience. Hige and colleagues have revealed how this may work in fruit flies. Does higher olfactory processing in vertebrates follow a similar scheme? A major target of the olfactory bulb (the vertebrate equivalent of the antennal lobe) is the piriform cortex, in which odours are represented by distributed activity patterns across many neurons. Unlike in the mushroom body, however, the output of the piriform cortex is not funnelled through a small neuronal population, suggesting that higher olfactory processing in vertebrates is more complex than in fruit flies. This notion is consistent with the richness of our olfactory perception and its powerful links to specific memories that we experience every day. ■

Thomas Frank and Rainer W. Friedrich
are at the Friedrich Miescher Institute for Biomedical Research, Basel 4058, Switzerland. R.W.F. is also at the University of Basel, Basel 4003, Switzerland.
e-mail: rainer.friedrich@fmi.ch

1. Laurent, G. *Nature Rev. Neurosci.* **3**, 884–895 (2002).
2. Hige, T., Aso, Y., Rubin, G. M. & Turner, G. C. *Nature* **526**, 258–262 (2015).
3. Murthy, M., Fiete, I. & Laurent, G. *Neuron* **59**, 1009–1023 (2008).
4. Caron, S. J. C., Ruta, V., Abbott, L. F. & Axel, R. *Nature* **497**, 113–117 (2013).
5. Waddell, S. *Curr. Opin. Neurobiol.* **23**, 324–329 (2013).
6. Séjourné, J. *et al. Nature Neurosci.* **14**, 903–910 (2011).
7. Cassenaer, S. & Laurent, G. *Nature* **482**, 47–52 (2012).
8. Aso, Y. *et al. eLife* **3**, e04577 (2014).
9. Gupta, N. & Stopfer, M. *Curr. Biol.* **24**, 2247–2256 (2014).
10. Aso, Y. *et al. eLife* **3**, e04580 (2014).
11. Proust, M. *Remembrance of Things Past. Vol. 1: Swann’s Way* (Vintage, 1913).

This article was published online on 30 September 2015.

Antiviral action countered by Nef

The HIV protein Nef is a viral ‘Swiss army knife’ with many functions. New work now shows how Nef increases infectivity — by inhibiting two of the host cell’s antiviral proteins, SERINC3 and SERINC5. [SEE ARTICLES P.212 & P.218](#)

CHRISTOPHER AIKEN

Nef is a small protein of the HIV virus that performs several diverse tasks during infection. It reduces the expression of a variety of proteins on the surface of the infected cell (usually, T cells of the host’s immune system), modulates T-cell signalling pathways, and increases the infectivity of new virus particles released from the cell. But for more than two decades, the mechanism by which Nef achieves this last function has been poorly understood. Now, in two breakthrough studies in this issue, Rosa *et al.*¹ (page 212) and Usami *et al.*² (page 218) show that Nef prevents the action of two host proteins previously not known to have antiviral activity: SERINC3 and SERINC5.

Work in the early 1990s showed that the ability of Nef to enhance viral infectivity is distinct from its well-characterized ability to downregulate the HIV receptor CD4, the removal of which is also important for infectivity. Although numerous studies had failed to find a strong effect of Nef on the structure or composition of HIV particles, one revealed that Nef had a mild enhancing effect on virus fusion with cells³. Other early studies showed that substituting the envelope proteins of HIV viruses, which mediate HIV entry into cells, with those of unrelated viruses can relieve the need for Nef in infection^{4,5}. The effect of such envelope swapping — known as pseudotyping — was specific, because only some viral proteins relieved the requirement for Nef. An unusual study that used mixed viral particles also linked the requirement for Nef with the viral-envelope proteins⁶. Yet Nef’s mechanism of enhancing infectivity remained elusive.

The Pizzato research group (who present the new paper by Rosa *et al.*) later uncovered two pieces of the puzzle, showing that Nef’s action depends on dynamin, a host protein that is crucial for the internalization of cell-surface proteins⁷, and that a glycosylated (carbohydrate-modified) variant of a structural protein from an unrelated retrovirus (the murine leukaemia virus (MLV) glyco-Gag protein) can also enhance HIV infectivity in a manner similar to that of Nef⁸. Although these clues did not reveal Nef’s mechanism,

they hinted that Nef had an effect on the cell, rather than on the virus particle. This may have pointed the investigators back to Nef’s ability to modulate the levels of cell-surface proteins. The researchers behind the second of the two new papers (Usami *et al.*) subsequently showed⁹ that HIV’s dependence on Nef varies between strains of the virus and is linked to a specific domain in the HIV envelope protein gp120.

In the current studies, the two groups independently pursued different approaches that converged on the same answer: incorporation of the host proteins SERINC3 and SERINC5 into HIV particles reduces infectivity through a mechanism that is countered by Nef (Fig. 1). Rosa *et al.* quantified global gene expression in a panel of cell lines in which HIV exhibits either high or low dependence on Nef, and noted a strong correlation between Nef dependence and SERINC5 expression. By contrast, Usami and co-workers compared the protein composition

of particles of normal HIV and Nef-defective HIV viruses that were released from cells in which virus infectivity strongly depends on Nef. This identified SERINC3 as a host protein that is enriched in particles that lack Nef.

Both SERINC3 and SERINC5 are members of a family of proteins named for a putative activity on membranes (‘serine incorporator’). The SERINC proteins are integral membrane proteins that form scaffolds for enzymes involved in the synthesis of specific membrane phospholipid molecules¹⁰. The two research groups investigated all five family members, but found that SERINC3 and SERINC5 are the only ones to induce antiviral activity against HIV. Both Nef and MLV glyco-Gag were shown to counter the antiviral action of SERINC3 and SERINC5. The two SERINC proteins are incorporated into Nef-defective HIV particles, but Nef and MLV glyco-Gag stop this from happening. They both induce SERINC5 to move from the cell surface to an intracellular compartment, preventing it from being incorporated into the budding virus, and demonstrating a plausible mechanism for how Nef enhances HIV infectivity.

The two groups also sought to determine whether pseudotyping HIV with Nef-independent surface proteins prevents the incorporation of SERINC3 or SERINC5 into Nef-defective HIV particles, but their different results leave this question open. However, Usami and colleagues demonstrated that the antiviral activity of SERINC3 and SERINC5 is specific for HIV surface proteins that cause virus infectivity to depend on Nef. Finally,

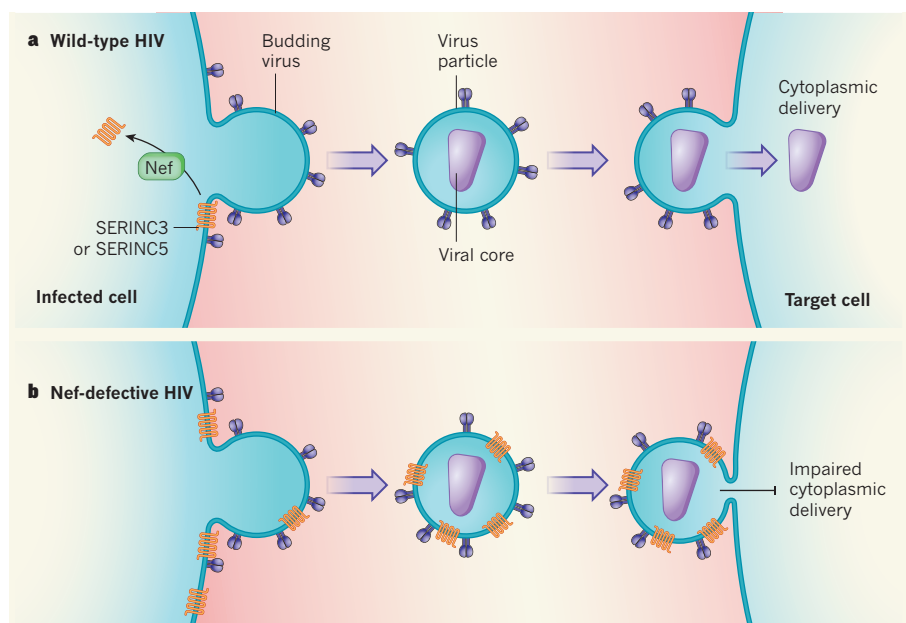


Figure 1 | SERINC proteins impair viral delivery. SERINC3 and SERINC5 are membrane proteins found by Rosa *et al.*¹ and Usami *et al.*² to have antiviral activity against HIV. **a**, The authors show that the HIV protein Nef prevents the SERINC proteins from being incorporated into a growing virus particle as it buds from the membrane of an infected cell. The resulting virus particle is able to correctly fuse with another target cell and deliver its viral core to the host-cell cytoplasm. **b**, The researchers propose that, in the absence of Nef, SERINC3 and SERINC5 are successfully incorporated into viral particles, and prevent delivery of the viral core by inhibiting the expansion of the fusion pore.

CONDENSED-MATTER PHYSICS

depletion of SERINC3 or SERINC5 from host cells selectively enhanced the infectivity of Nef-defective HIV, confirming that SERINC proteins reduce HIV infectivity. Both SERINC3 and SERINC5 are expressed in human blood cells, suggesting that they are active in the primary targets of HIV infection *in vivo*.

How do SERINC3 and SERINC5 lower HIV infectivity? Both groups showed that, at high levels of expression, these proteins reduce the efficiency of virus fusion with target cells. However, the proteins inhibited HIV infectivity more strongly than they affected fusion, suggesting that they also affect an early post-fusion step in infection. Accordingly, Rosa *et al.* propose a model in which SERINC3 and SERINC5 prevent the expansion of the pore that is formed between the viral and cell membranes and that is necessary for delivery of the viral core into the cell's cytoplasm (Fig. 1). However, it is not clear whether this would result in the impaired reverse transcription of the viral genome that is normally seen in Nef-defective HIV. One intriguing possibility is that the impeded viral core is targeted for cellular destruction, as previously suggested¹¹.

The identification of SERINC3 and SERINC5 as antiviral proteins that are counteracted by diverse retroviruses suggests that these proteins may also target other enveloped viruses, which may in turn have different mechanisms for escaping their antiviral action. These two proteins can therefore be used as probes to examine the entry mechanisms of enveloped viruses. Although the available data suggest that these proteins target specific regions of viral glycoproteins, it is possible that they inhibit fusion indirectly by controlling the lipid composition or fluidity of the viral membrane. Regardless of the specific antiviral mechanism, the ability of Nef to counteract SERINC3 and SERINC5 adds to the impressive list of functions of this remarkable little viral protein. ■

Christopher Aiken is in the Department of Pathology, Microbiology and Immunology, Vanderbilt University School of Medicine, Nashville, Tennessee 37232-2363, USA. e-mail: chris.aiken@vanderbilt.edu

- Rosa, A. *et al.* *Nature* **526**, 212–217 (2015).
- Usami, Y., Wu, Y. & Göttinger, H. G. *Nature* **526**, 218–223 (2015).
- Day, J. R., Münk, C. & Guatelli, J. C. *J. Virol.* **78**, 1069–1079 (2004).
- Aiken, C. J. *Virol.* **71**, 5871–5877 (1997).
- Chazal, N., Singer, G., Aiken, C., Hammarskjöld, M.-L. & Rekosh, D. *J. Virol.* **75**, 4014–4018 (2001).
- Zhou, J. & Aiken, C. J. *Virol.* **75**, 5851–5859 (2001).
- Pizzato, M. *et al.* *Proc. Natl Acad. Sci. USA* **104**, 6812–6817 (2007).
- Pizzato, M. *Proc. Natl Acad. Sci. USA* **107**, 9364–9369 (2010).
- Usami, Y. & Göttinger, H. *Cell Rep.* **5**, 802–812 (2013).
- Inuzuka, M., Hayakawa, M. & Ingi, T. *J. Biol. Chem.* **280**, 35776–35783 (2005).
- Qi, M. & Aiken, C. J. *Virol.* **81**, 1534–1536 (2007).

This article was published online on 30 September 2015.

Quantum dots and the Kondo effect

Nanotechnology studies explore the extreme properties of strongly interacting electronic systems through conductance measurements, and probe quantum phase transitions close to absolute zero temperature. [SEE LETTERS P.233 & P.237](#)

KARYN LE HUR

Magnetic impurities, in the form of atoms that have spin angular momentum (spin) of $\frac{1}{2}$, can drastically modify the electrical resistivity of metals. In the presence of such atoms, the resistivity drops to a minimum at a certain low temperature a few kelvin above absolute zero, but then increases as the temperature falls further^{1,2}. This behaviour at low temperatures is called the Kondo effect, and it has been modelled in nanosystems known as quantum dots (QDs) that have been engineered to behave as spin- $\frac{1}{2}$ artificial atoms³.

To model the Kondo effect, a QD is connected by means of quantum tunnelling to two electron reservoirs through electrodes that form electron-transport channels — a set-up known as the one-channel Kondo model. Writing on pages 233 and 237 of this issue, respectively, Iftikhar *et al.*⁴ and Keller *et al.*⁵ report on experiments with QDs based on gallium–aluminium–arsenic interfaces, in which they implement an improved set-up called the two-channel Kondo model⁶. Keller *et al.* go on to test fundamental properties of quantum phase transitions in their systems near absolute zero temperature⁷.

In metals, the mechanism responsible for

the Kondo effect at low temperatures is the coupling between a magnetic impurity and the spins of conduction electrons. At temperatures lower than the Kondo temperature T_K , the electrons interact strongly with the impurity (a process known as screening), forming a collection of heavy quasiparticles^{1,2} called a Fermi liquid. In the one-channel QD set-up, electrons in the reservoirs enter a state of superposition that forms a single electron channel, which strongly couples with the QD's spin near the Fermi level (the energy of the highest filled electronic level of a system at absolute zero temperature). The conductance through the channel reaches a maximum value below T_K (which is much lower than 1 kelvin in QDs³), in contrast to the increased resistivity in metals with magnetic impurities.

The two-channel Kondo model involves two sources of electrons that form two separate electron channels which compete to screen the impurity, thereby producing non-trivial low-temperature effects predicted by Kondo theory⁸. Iftikhar and colleagues observe the two-channel effect using a micrometre-scale QD (Fig. 1a) that comprises an 'island' of several billion electrons⁴. On this scale, the QD behaves like a metal, but its charge is quantized; this means that, to add one electron to the QD, a finite amount of charging energy

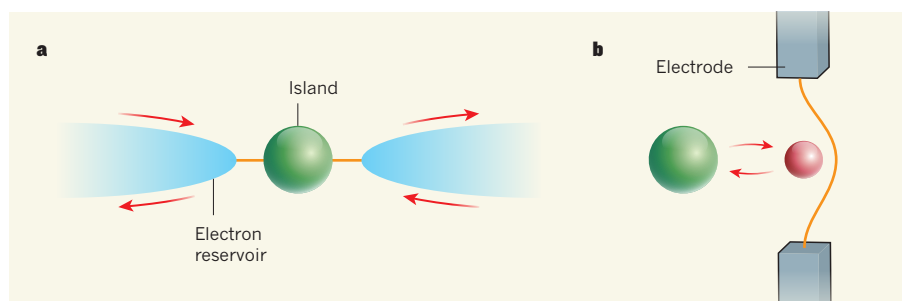


Figure 1 | Strongly interacting electronic systems. **a**, Iftikhar *et al.*⁴ model the one- and two-channel Kondo effect using a device consisting of a micrometre-scale semiconductor known as a quantum dot, or island, that has a pseudo-spin angular momentum $\frac{1}{2}$ and is connected to two electron reservoirs through nanometre-scale constrictions (orange) called quantum point contacts (QPCs). The authors place the device perpendicular to a strong magnetic field and measure how the conductance at each QPC changes at extremely low temperatures as single electrons enter or exit the quantum dot (red arrows). **b**, Keller and colleagues' device⁵, which probes the difference between the one- and two-channel effects, consists of a nanometre-scale spin- $\frac{1}{2}$ quantum dot (red) that is coupled (red arrows) to a metallic electron island, forming one electron-transport channel, and to source and drain electrodes, which together form a separate channel. The authors measure the conductance between the electrodes and investigate transitions between different quantum phases of matter at temperatures close to absolute zero. (**b** Adapted from ref. 5.)

is required. The QD's $\frac{1}{2}$ spin results from a resonance between two macroscopic charge states that have the same energy⁹. When electrons enter or exit the island, the QD's spin changes direction.

The authors connect the QD to external electron reservoirs through two macroscopic electrodes and apply a strong magnetic field (3.9 tesla). The field polarizes the spins of individual electrons, which restricts the number of quantum electron channels to two. The authors demonstrate the two-channel effect through two quantum point contacts (QPCs), which are nanoconstrictions separating the QD from the electron reservoirs.

Iftikhar *et al.* obtain a first indication that two-channel Kondo physics is at work in their system from an analysis of conductance measurements, which yield T_K as a function of the transmission probability of a single electron through the QPCs. The increase of the conductance below T_K that they observe is in agreement with Kondo theory¹⁰. At each of the two QPCs, the conductance shows a maximum value of e^2/h , where e is the electron charge and h the Planck constant. Taking both QPCs into account, the conductance through the system becomes $e^2/(2h)$, by analogy to two resistors connected in series¹⁰.

The authors then observe how the conductance changes with temperature, and expose the fragility of the two-channel effect: they find that this effect occurs only for a narrow range of conductances and temperatures; outside that range, it transitions to the one-channel effect, in agreement with theory^{10,11}. The channel that couples most strongly to the QD screens the QD at the low-temperature limit (14 millikelvin). An open question is why the conductance of the most strongly coupled QPC slightly exceeds e^2/h — the quantum upper limit.

Keller and colleagues explore the difference at the quantum level in the conductance between the one- and two-channel effects using highly controllable nanotechnology⁵ and theoretical and numerical arguments. The authors fabricate a quantum device¹² consisting of a nanometre-scale spin- $\frac{1}{2}$ QD that has an odd number of electrons³ and two electron channels: a pair of source and drain electrodes together form one 'delocalized' channel, and a metallic electron island acts as a separate channel (Fig. 1b). In contrast to Iftikhar and colleagues' system, the number of electrons in the metallic island cannot be modified in this experiment.

The microscopic origins of transitions between different quantum phases of matter at zero temperature are not always understood and are often debated. According to the theory of the two-channel Kondo effect⁸, the temperature (or conductance) domain close to the quantum phase transition is well understood and can be studied not only by tuning the strength of the coupling of each (macroscopic) quantum channel with the

spin- $\frac{1}{2}$ impurity, but also by adjusting external parameters such as the temperature, the magnetic field or the bias and gate voltages of nanosystems. This leads to a complex behaviour in the electron-transport properties across the quantum phase transition. When the two channels couple equally to the spin- $\frac{1}{2}$ QD, the electron transport for temperatures below T_K (corresponding to thermal energies less than about 50 microelectronvolts) cannot be interpreted by the standard quasiparticle picture, leading to the emergence of non-Fermi-liquid behaviour.

In their system, Keller *et al.* observe that, when the spins of the two electron channels couple to the nano-QD in such a way that the energy exchange between the channels and the QD (which acts as the magnetic impurity) is different in the two cases, the more strongly coupled channel pairs with the QD. The conductance measurements then show that the system undergoes a quantum phase transition, reverting to a Fermi liquid of quasiparticles (whose wavefunction acquires a 90° phase shift). At a temperature above absolute zero, the smooth transition between the non-Fermi-liquid and Fermi-liquid regimes involves a distinct energy scale (T^*), and the authors' conductance measurements are in agreement with precise theoretical predictions of that energy scale^{13,14}. Keller *et al.* report a quadratic dependence of T^* on the gate voltage, and confirm the exact theoretical description of the transition between the strongly correlated non-Fermi-liquid and Fermi-liquid states.

Keller and co-workers' device essentially

provides a sophisticated 'nanoscope' with which the authors explore domains close to quantum phase transitions in an extremely narrow physical-parameter space. Iftikhar and colleagues' impressive experiment demonstrates the two-channel Kondo effect on the basis of the quantum charge states of a microscopic QD in an electric circuit. Follow-up research could involve more electron channels by increasing the numbers of QPCs in an effort to investigate other similar systems that are best described in terms of strongly interacting entities (strongly correlated physics). ■

Karyn Le Hur is at the Centre for Theoretical Physics (CPHT) at the Ecole Polytechnique, and at CNRS, Université Paris-Saclay, 91128 Palaiseau, France.

e-mail: karyn.le-hur@polytechnique.edu

1. Kondo, J. *Prog. Theor. Phys.* **32**, 37–49 (1964).
2. Nozières, P. *J. Low Temp. Phys.* **17**, 31–42 (1974).
3. Kouwenhoven, L. & Glazman, L. *Phys. World* **14**(1), 33–38 (2001).
4. Iftikhar, Z. *et al. Nature* **526**, 233–236 (2015).
5. Keller, A. J. *et al. Nature* **526**, 237–240 (2015).
6. Nozières, P. & Blandin, A. *J. Phys.* **41**, 193–211 (1980).
7. Sachdev, S. *Quantum Phase Transitions* 2nd edn (Cambridge Univ. Press, 2011).
8. Cox, D. L. & Zawadowski, A. *Adv. Phys.* **47**, 599–942 (1998).
9. Matveev, K. A. *Sov. Phys. JETP* **72**, 892–899 (1991).
10. Furusaki, A. K. & Matveev, A. *Phys. Rev. B* **52**, 16676 (1995).
11. Le Hur, K. & Seelig, G. *Phys. Rev. B* **65**, 165338 (2002).
12. Potok, R., Rau, I. G., Shtrikman, H., Oreg, Y. & Goldhaber-Gordon, D. *Nature* **446**, 167–171 (2007).
13. Affleck, I. & Ludwig, A. W. W. *Phys. Rev. B* **48**, 7297–7321 (1993).
14. Sela, E., Mitchell, A. K. & Fritz, L. *Phys. Rev. Lett.* **106**, 147202 (2011).

ASTROPHYSICS

Surprisingly fast motions in a dust disk

A recently commissioned planet-finding instrument has been used to study a young solar system around the star AU Microscopii, leading to the discovery of rapidly moving features in the dust disk around the star. SEE LETTER P.230

MARSHALL D. PERRIN

In the southern sky hangs the constellation Microscopium, 'the microscope', one of several minor constellations named after scientific instruments in the eighteenth century. Using some much more recent instrumentation, Boccaletti *et al.*¹ (page 230 of this issue) have now observed with fresh clarity a young solar system nestled within that constellation, homing in on a dusty ring of debris around the star AU Microscopii (AU Mic). These observations from the ground match or exceed the resolution with which this debris

disk was previously seen by the Hubble Space Telescope. When the authors compared the new images with the older Hubble data, they discovered several localized areas of enhanced brightness, perhaps clouds or clumps of dust, moving outwards from the star surprisingly fast on trajectories suggesting that the clouds are likely to escape into interstellar space.

Twenty years ago this month, astronomers announced the detection of a planet orbiting a Sun-like star², launching a revolution that has led to detections of thousands of exoplanets with a dizzying diversity of properties. But the vast majority of such discoveries have been

made indirectly, primarily by measuring minute variations in the light of the host star from which the presence of a planet can be inferred.

Directly imaging anything orbiting a nearby star is a daunting observational challenge; even large planets such as Jupiter are hundreds of thousands or millions of times fainter than the associated star, and are easily lost in the glare of scattered starlight. But that challenge is worth pursuing, because if one can see something directly, one can measure its spectrum, enabling detailed physical characterization. The same holds true for rings of dusty debris, which can be produced by the evaporation of comets or collisional destruction of asteroids. All-sky infrared surveys have taught us that such debris disks are common, but only a small fraction has been seen directly.

AU Mic hosts one such disk, first imaged more than a decade ago³. As stars go, it is small, nearby and young: half the mass of the Sun, 10 parsecs (32 light years) away and about 25 million years old. Observations⁴ have established that it is surrounded by a belt of planetesimals at a radius of about 40 astronomical units (1 AU is Earth's distance from the Sun), similar to the Kuiper belt of our Solar System. Occasional collisions of the planetesimals in the AU Mic disk liberate dust particles that are blown outwards by the stellar wind. The total mass of dust is about that of the Moon, ground fine and spread widely. And just as our Solar System's dust is non-uniformly distributed, the disk around AU Mic has clumps and asymmetries, which some have interpreted^{5–8} as signs of an unseen planet stirring up the smaller bodies.

This intriguing evidence led to AU Mic becoming one of the first targets for SPHERE (Spectro-Polarimetric High-contrast Exoplanet Research; Fig. 1), a planet-finding instrument, soon after it started operating in 2014. SPHERE is a specialized 'high-contrast' imaging system, designed to allow the Very Large Telescope in Chile to point towards a bright star while blocking almost all of that star's light, and thus allowing a view of the star's close environs. Developed over the past decade, it uses a combination of sophisticated optical, spectroscopic and analytical techniques, and was intended to enable the detection of light up to one million times fainter than the host star.

So, did SPHERE meet expectations? Boccaletti *et al.* present a convincing affirmative with their observations of AU Mic. Their data are beautiful and impressive, matching the Hubble images in terms of clarity of fine structure, and exceeding Hubble in terms of the ability to peer close to the central star. What was definitely not expected, however, was what they saw around AU Mic: the clumpy dust clouds that extend across much of the disk to the southeast of the star have moved outwards over the past few years, by between 3 and 8 AU each (see Fig. 1 of the paper¹).

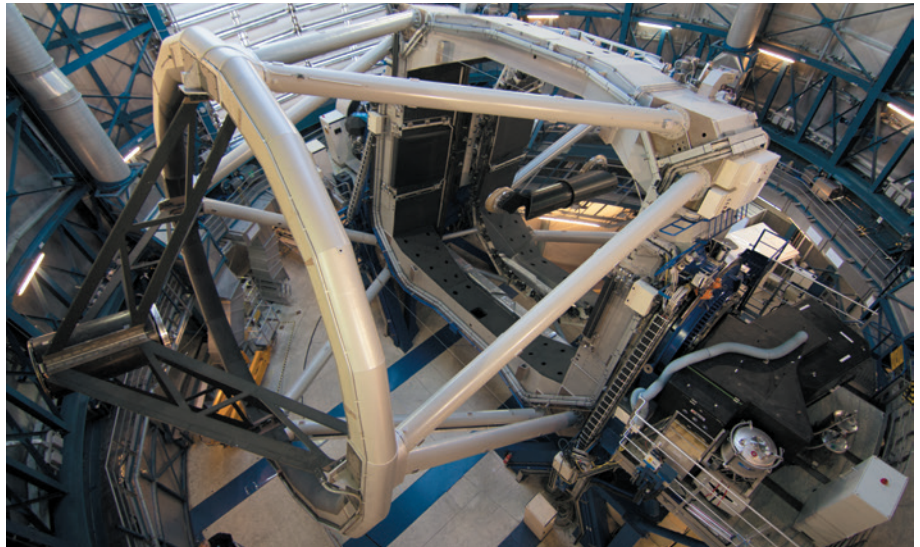


Figure 1 | Spectro-Polarimetric High-contrast Exoplanet Research (SPHERE). The imaging system SPHERE has been attached to the Very Large Telescope in Chile (SPHERE is the black box in the bottom right of the photo). Boccaletti *et al.*¹ have used it to image the dust disk around the star AU Microscopii.

This unexpected result led the team to revisit some of the earlier Hubble data with more careful analysis. By separately analysing two data sets that had previously been combined, they found that the outward motion could also be seen in the Hubble data between 2010 and 2011. Although it is well understood that individual dust grains would be gradually blown out of the system, this would be expected to be a fairly continuous, steady-state process — not something that would produce a chain of discrete clouds of dust, stretching across an apparent distance equal to the diameter of our Solar System and moving outward as a coherent pattern. Furthermore, the observed speed increases roughly linearly with apparent distance from the star, from 4 to 10 kilometres per second; the motion of the outermost clumps is fast enough for bound orbits around the star to be ruled out. These clouds seem to be blowing out into interstellar space.

The authors readily admit that they do not have a good explanation for what is going on. They present several potential hypotheses, from resonant waves of dust induced by an unseen planet, to debris from massive asteroid collisions, to material ejected from the debris ring as a result of periodic stellar flares. But none of these is fully satisfactory. One highly speculative idea is that stellar flares are interacting with a planetary magnetosphere (the region in which charged particles are affected by a planet's magnetic field) or a circumplanetary ring like that of Saturn, in which case the projected orbital motion of a planet around the star could explain the variation in cloud velocities. Further investigations of AU Mic's debris disk are surely a high priority for SPHERE, even as the instrument moves into full use studying a large sample of targets.

SPHERE is not the only game in town for high-contrast imaging. Over the past few

years, instruments of this type have become available at many large telescopes, each with their own strengths and specializations. In fact, the AU Mic system has also recently been observed by the Gemini Planet Imager at the Gemini South Telescope in Chile⁹, although this instrument's field of view is too small to see the high-speed outer clumps detected by SPHERE. Additional competition comes from several instruments that detect slightly longer infrared wavelengths (SPHERE operates in the visible and near-infrared regions of the spectrum), at which planets can be brighter and the demands for adaptive-optics systems are less stringent.

But even with the latest instrumentation and large surveys, the challenge of imaging planets is so great that only a modest number are likely to be seen over the next few years. Cases such as that of AU Mic, in which disks can be imaged in great detail but any planets present are unseen, are likely to remain more common than directly imaged planets. Lucky for astronomers, then, that circumstellar disks still turn out to have surprises such as the fast-moving dust features of AU Mic. ■

Marshall D. Perrin is at the *Space Telescope Science Institute, Baltimore, Maryland 21218, USA.*
e-mail: mperrin@stsci.edu

1. Boccaletti, A. *et al. Nature* **526**, 230–232 (2015).
2. Mayor, M. & Queloz, D. *Nature* **378**, 355–359 (1995).
3. Kalas, P., Liu, M. C. & Matthews, B. C. *Science* **303**, 1990–1992 (2004).
4. Strubbe, L. E. & Chiang, E. I. *Astrophys. J.* **648**, 652–665 (2006).
5. Liu, M. C. *Science* **305**, 1442–1444 (2004).
6. Metchev, S. A., Eisner, J. A., Hillenbrand, L. A. & Wolf, S. *Astrophys. J.* **622**, 451–462 (2005).
7. Krist, J. E. *et al. Astron. J.* **129**, 1008–1017 (2005).
8. Fitzgerald, M. P., Kalas, P. G., Duchêne, G., Pinte, C. & Graham, J. R. *Astrophys. J.* **670**, 536–556 (2007).
9. Wang, J. J. *et al. Astrophys. J. Lett.* (in the press); preprint at <http://arxiv.org/abs/1508.04765> (2015).

The effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015

S. Bhatt^{1*}, D. J. Weiss^{1*}, E. Cameron^{1*}, D. Bisanzio¹, B. Mappin¹, U. Dalrymple¹, K. E. Battle¹, C. L. Moyes¹, A. Henry¹, P. A. Eckhoff², E. A. Wenger², O. Briët^{3,4}, M. A. Penny^{3,4}, T. A. Smith^{3,4}, A. Bennett⁵, J. Yukich⁶, T. P. Eisele⁶, J. T. Griffin⁷, C. A. Fergus⁸, M. Lynch⁸, F. Lindgren⁹, J. M. Cohen¹⁰, C. L. J. Murray¹¹, D. L. Smith^{1,11,12,13}, S. I. Hay^{11,13,14}, R. E. Cibulskis⁸ & P. W. Gething¹

Since the year 2000, a concerted campaign against malaria has led to unprecedented levels of intervention coverage across sub-Saharan Africa. Understanding the effect of this control effort is vital to inform future control planning. However, the effect of malaria interventions across the varied epidemiological settings of Africa remains poorly understood owing to the absence of reliable surveillance data and the simplistic approaches underlying current disease estimates. Here we link a large database of malaria field surveys with detailed reconstructions of changing intervention coverage to directly evaluate trends from 2000 to 2015, and quantify the attributable effect of malaria disease control efforts. We found that *Plasmodium falciparum* infection prevalence in endemic Africa halved and the incidence of clinical disease fell by 40% between 2000 and 2015. We estimate that interventions have averted 663 (542–753 credible interval) million clinical cases since 2000. Insecticide-treated nets, the most widespread intervention, were by far the largest contributor (68% of cases averted). Although still below target levels, current malaria interventions have substantially reduced malaria disease incidence across the continent. Increasing access to these interventions, and maintaining their effectiveness in the face of insecticide and drug resistance, should form a cornerstone of post-2015 control strategies.

In the midst of an escalating malaria public health disaster, the year 2000 marked a turning point in multilateral commitment to malaria control in sub-Saharan Africa, catalysed by the Roll Back Malaria initiative and the wider development agenda around the United Nations Millennium Development Goals (MDGs). The 15 years since have seen international financing for malaria control increase approximately twentyfold¹, enabling widespread but uneven scale-up of coverage of the main contemporary malaria control interventions: insecticide-treated bed nets (ITNs), indoor residual spraying (IRS), and prompt treatment of clinical malaria cases with artemisinin-based combination therapy (ACT).

As part of this reinvigorated effort, a series of international goals were set with a target year of 2015, in particular the MDG to “halt by 2015 and begin to reverse the incidence of malaria” and the more ambitious target defined later by the World Health Organization (WHO) of reducing case incidence by 75% relative to 2000 levels². While these targets were important for motivating action and mobilizing funds, no explicit plan was put in place to reliably measure progress towards them. Now that the benchmark year of 2015 has been reached, the international community must define a post-2015 agenda for malaria control that will shape the technical, financial and political landscape in which the battle against the disease will be

fought. This agenda is being defined around two key policy initiatives for the 2016–2030 period: the Global Technical Strategy³ and Action and Investment to Defeat Malaria⁴, led by WHO and the Roll Back Malaria Partnership. In this context, it is imperative that the achievements of 2015 can be robustly evaluated and, more broadly, that the patterns, causes, and implications of changing malaria endemicity over the past 15 years can be understood to inform an optimal strategy for the future.

The effect of malaria control is poorly understood

Despite its importance, current knowledge on the nature and drivers of changing endemicity in sub-Saharan Africa is remarkably weak. National health records in 32 highly endemic countries (together accounting for about 90% of the global malaria burden) are considered inadequate to assess trends in malaria cases¹. This stems from low care-seeking rates (many malaria cases are not seen at formal health facilities), incomplete record keeping and curation (many recorded cases are never captured in surveillance databases), and historically poor access to parasitological diagnosis (malaria cases were often diagnosed presumptively with poor specificity). As systems have begun to improve, for example owing to greater use of rapid diagnostic testing in health facilities¹, these biases have been mitigated,

¹Spatial Ecology and Epidemiology Group, Tinbergen Building, Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK. ²Institute for Disease Modeling, Intellectual Ventures, 1555 132nd Avenue NE, Bellevue, Washington 98005, USA. ³Epidemiology and Public Health, Swiss Tropical and Public Health Institute, 4002 Basel, Switzerland. ⁴University of Basel, Petersplatz 1, 4001 Basel, Switzerland. ⁵Malaria Elimination Initiative, University of California San Francisco, 500 Parnassus Avenue, San Francisco, California 94143, USA. ⁶Center for Applied Malaria Research and Evaluation, Tulane University School of Public Health and Tropical Medicine, 1440 Canal Street, Suite 2200 New Orleans, Louisiana 70112, USA. ⁷MRC Centre for Outbreak Analysis and Modelling, Department of Infectious Disease Epidemiology, Imperial College London, London W2 1PG, UK. ⁸Global Malaria Programme, World Health Organization, 20 Avenue Appia, 1211 Geneva 27, Switzerland. ⁹Department of Mathematical Sciences, University of Bath, Claverton Down, Bath BA2 7AY, UK. ¹⁰Clinton Health Access Initiative, Boston, Massachusetts 02127, USA. ¹¹Institute for Health Metrics and Evaluation, University of Washington, Seattle, Washington 98121, USA. ¹²Sanaria Institute for Global Health and Tropical Medicine, Rockville, Maryland 20850, USA. ¹³Fogarty International Center, National Institutes of Health, Bethesda, Maryland 20892-2220, USA. ¹⁴Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK.

*These authors contributed equally to this work.

but that presents further challenges for comparison of data through time and evaluation of trends⁵. For these countries, the WHO has previously adopted a “cartographic” burden measurement approach whereby a map of climatic suitability for malaria transmission is first used to stratify likely incidence rates across the continent, with these rates then progressively downgraded as intervention coverage increases according to effect sizes measured in randomized control trials. One acknowledged limitation of this approach is its reliance on the central assumption that effects observed in a limited number of short-term trials can be extrapolated to sustained continent-wide implementation. This assumption has never been validated beyond local or national-level analyses⁶. In reality, the individual or combined efficacy of interventions will vary by setting and be contingent on many local factors, including vector ecology, health systems, and coverage levels^{7,8}. Other studies have investigated effects using cross-sectional community surveys^{7,9}. These studies capture a wider range of real-world settings, but yield only a single pooled estimate across diverse disease transmission settings that, again, has unknown validity when extrapolated across Africa.

Since its first use, component parts of the cartographic burden framework have incrementally improved. Climatic suitability maps have been superseded by empirical endemicity maps^{10–12} that use model-based geostatistics to create surfaces of risk based on thousands of geolocated cross-sectional surveys measuring infection prevalence (termed *Plasmodium falciparum* parasite rate, *PfPR*). Improvements have also been made in the estimation of clinical incidence rates as a function of *PfPR*^{13–15}, allowing clinical incidence rates, which are notoriously difficult to measure in the field, to be estimated geographically using mapped surfaces of *PfPR*^{14,16}. All these earlier studies, however, preceded the most intense period of control effort (from 2010 to the present), and none were designed to formally evaluate temporal changes in disease burden or explicitly consider the effect of interventions.

A framework to measure malaria risk in Africa

Here, we provide the first formal quantification, with rigorously defined uncertainty, of *P. falciparum* infection prevalence and disease incidence across sub-Saharan Africa from the year 2000 to the benchmark year of 2015, and of the role the major control interventions have had in causing these changes. Our approach evaluates not only point estimates, but also presents a full treatment of uncertainty through a Bayesian hierarchical model. Components contributing to uncertainty in outputs included the sample size and spatiotemporal density of *PfPR* surveys, uncertainty in the fitted relationships between *PfPR* and the suite of environmental and intervention covariates, uncertainty in the input data on observed clinical incidence rates, and uncertainty in the mechanistic model parameters defining the prevalence–incidence relationship. By linking all components together in a Bayesian framework, these distinct sources of uncertainty are formally propagated through the predictive model and represented as predictive posterior distributions around all output results.

The analytical framework is shown schematically in Extended Data Fig. 1. Data on ITN use and access to ACTs from over one million households were combined with national malaria control programme data¹ on ITN, ACT and IRS provision to develop time-series models of coverage of these interventions within each country¹. These were combined within a spatiotemporal Bayesian geostatistical model¹⁷ with *PfPR* data from 27,573 georeferenced population clusters between 1995 and 2014, along with an optimised suite of temporally dynamic environmental and sociodemographic covariates¹⁸. The model adjusted *PfPR* observations by age¹⁹, season and type of diagnostic used, and fitted flexible functional forms to capture the effect of each intervention on declining *PfPR* as a function of coverage reached and the starting (pre-intervention) *PfPR* in 2000 (Extended Data Fig. 2). Following earlier work^{10,11,19}, we chose to model *PfPR* in

the 2-up-to-10 year age range, since this is associated with a plateau in the age-prevalence relationship and thus acts as a standardised comparison. The model was used to predict a spatio-temporal ‘cube’ of age-structured *PfPR* at 5×5 km resolution across all endemic African countries for each year from 2000 to 2015. Using the empirically observed effect of each intervention, it was possible to generate counterfactual maps estimating contemporary *PfPR* under hypothetical scenarios without interventions. We chose to evaluate this ‘no intervention’ counterfactual to allow estimation of the total effect of interventions.

For the 32 high-burden countries of Africa, an ensemble model was developed to predict incidence rates of clinical malaria as a function of community *PfPR*²⁰. This brought together three independently developed mathematical malaria transmission models^{14,15,21} that were re-fitted to a common data set of age-structured clinical incidence measured longitudinally at 30 sites²², allowing an ensemble model to be defined to predict age-specific incidence at all locations given prevalence, seasonality, level of treatment, and probable immune status of populations. We used a definition of ‘clinical malaria’ as an attributable febrile episode (body temperature in excess of 37.5 °C), censored by a 30-day window (that is, multiple bouts of symptoms occurring within the same 30-day period are counted as a single episode). The ensemble model was then combined with the *PfPR* cube and underlying population surfaces²³ to predict clinical incidence by country and year for both the real and counterfactual scenarios. For the remaining eleven low-burden countries in Africa (accounting for around 3% of cases) where national reporting systems are more robust, we generated clinical incidence estimates with an existing approach that uses national case reports while adjusting for care-seeking behaviour, low diagnostic testing rates, and underreporting^{1,24}.

Infection prevalence and clinical incidence decline

We found that infection prevalence in children age 2-up-to-10 across endemic Africa has halved since the year 2000 (population-weighted mean *PfPR*_{2–10}; year 2000 = 33%, 95% credible interval 31–35%; year 2015 = 16%, 14–19%), with around three-quarters of this decline occurring after 2005. Across Africa the rate of decline in *PfPR*_{2–10} rose steadily to a peak yearly decline of 9% in 2011, after which there was a slowing between 2011 and 2013 followed by resurgence in recent years back to the current rate of 5% annual decline in *PfPR*. Our predicted surfaces of *PfPR*_{2–10} demonstrate the geographical pattern of this reduction across the continent (Fig. 1a–c), with hyper- or holo-endemic transmission (where *PfPR* exceeds 50% and 75%, respectively, see Fig. 1d) common in 2000 across large swathes of central and western Africa, but limited to isolated pockets by 2015. This decline meant a marked shift in the distribution of exposure level (Fig. 1d), with the proportion of the endemic population exposed to hyper- or holo-endemic malaria falling from 33% (30–37%) to just 9% (5–13%) (Table 1). Crucially, for the feasibility of post-2015 elimination efforts, the population of stable endemic Africa experiencing very low transmission (*PfPR*_{2–10} less than 1%) has increased sixfold since 2000 (far outpacing the 50% underlying population growth over the period) meaning there are now 121 (110–133) million people living in settings where elimination campaigns can be considered.

We estimated that there were 187 (132–259) million clinical cases of *P. falciparum* malaria in Africa in 2015. Case incidence declined by 40% from 321 (253–427) per 1,000 persons per annum in 2000 to 192 (135–265) per 1,000 persons p.a. in 2015, with all but one of the 43 mainland endemic countries meeting the MDG target of reversing incidence trends by 2015, 19 (17–25) achieving a >50% decline, and 7 (6–7) declining by >75% (Extended Data Fig. 3).

The model has been able to predict changes in mean *PfPR* across Africa with considerable precision, reflecting the increasing abundance of *PfPR* surveys, their relatively large signal-to-noise ratio, the adjustments for diagnostic type, and the informative covariate suite, all of which contributed to strong predictive performance of the geostatistical

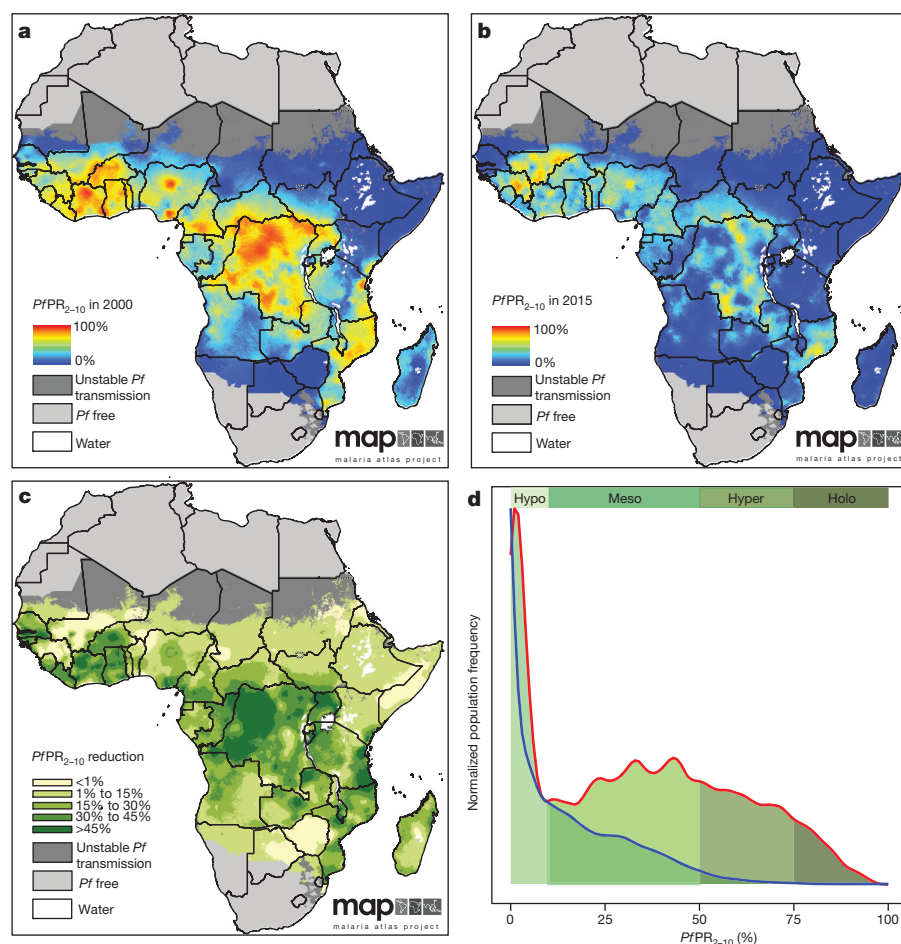


Figure 1 | Changes in infection prevalence 2000–2015. **a**, $PfPR_{2-10}$ for the year 2000 predicted at 5×5 km resolution. **b**, $PfPR_{2-10}$ for the year 2015 predicted at 5×5 km resolution. **c**, Absolute reduction in $PfPR_{2-10}$ from 2000 to 2015. **d**, Smoothed density plot showing the relative distribution of endemic populations by $PfPR_{2-10}$ in the years 2000 (red line) and 2015 (blue line). The frequencies on the vertical axis have been scaled to make the densities visually comparable. The classical endemicity categories are shown for reference in green shades. Results shown in all panels are derived from a Bayesian geostatistical model fitted to $n = 27,573$ $PfPR$ survey points; $n = 24,868$ ITN survey points; $n = 96$ national survey reports of ACT coverage; $n = 688$ country-year reports on ITN, ACT and IRS distribution by national programs; and $n = 20$ environmental and socioeconomic covariate grids. Maps in **a–c** are available from the Malaria Atlas Project (<http://www.map.ox.ac.uk/>) under the Creative Commons Attribution 3.0 Unported License.

model. Credible intervals around the continental clinical incidence estimates were proportionately much larger and this reflected primarily the residual uncertainty around the modelled relationship between infection prevalence and clinical incidence.

Attributable effect of malaria control interventions

Changes in prevalence largely followed patterns of increasing ITN coverage, and ITNs were by far the most important intervention across Africa, accounting for an estimated 68 (62–72)% of the declines in $PfPR$ seen by 2015 (Fig. 2a). We estimated ACT and IRS contributed 19 (15–24)% and 13 (11–16)% respectively, although these interventions had larger proportional contributions where their coverage was high (Extended Data Fig. 4). It is important to emphasize that these proportional contributions do not necessarily reflect the comparative effectiveness of different intervention strategies but, rather, are driven primarily by how early and at what scale the different interventions were deployed. In total, we estimated that malaria control interventions have averted 663 (542–753) million clinical cases since 2000, of which 68 (62–73)%, 22 (17–28)% and 10 (5–14)% were contributed by ITNs, ACTs, and IRS, respectively (Fig. 2b).

Discussion

Here, for the first time, the rapidly changing landscape of malaria risk in Africa has been quantified across the 15-year span of the Millennium Development Goals. Our approach is primarily data driven, informed by empirical observations in the field rather than theoretical models or extrapolated experimental results. Our modelling framework requires few prior assumptions and allows patterns of change and attribution to be identified with rigorously defined metrics of uncertainty.

We have shown that remarkable and widespread reductions in infection prevalence and case incidence have occurred across Africa since 2000, and that malaria control interventions have been responsible for most of the decline even though they remain well below international targets for universal coverage¹. ITNs have had by far the largest effect, but have also been generally present for longer and at higher levels of coverage. IRS and ACTs have both made important contributions to reducing prevalence and incidence where they have been implemented at scale (although it is important to note that the primary role of ACTs is in averting severe disease and death rather than reducing transmission and uncomplicated cases).

Table 1 | Changing distribution of malaria endemicity across stable endemic Africa, 2000 to 2015

| Endemicity class | Population (%) | | | Area (%) | | |
|--|----------------|--------|------------|----------|--------|------------|
| | 2000 | 2015 | Change (%) | 2000 | 2015 | Change (%) |
| Holo ($PfPR_{2-10} \geq 75\%$) | 11.57 | 1.32 | –88.57 | 11.81 | 1.38 | –88.32 |
| Hyper ($PfPR_{2-10}$ 50–75%) | 21.51 | 7.46 | –65.31 | 20.18 | 7.88 | –60.93 |
| Meso ($PfPR_{2-10}$ 10–50%) | 41.32 | 42.41 | +2.64 | 40.98 | 41.06 | +0.19 |
| Hypo ($PfPR_{2-10} < 10\%$) | 25.60 | 48.80 | +90.63 | 27.02 | 49.67 | +83.84 |
| Total | 100.00 | 100.00 | | 100.00 | 100.00 | |
| Pre-elimination or eliminating ($PfPR_{2-10} < 1\%$) | 3.82 | 13.61 | +255.93 | 3.48 | 11.39 | +227.51 |

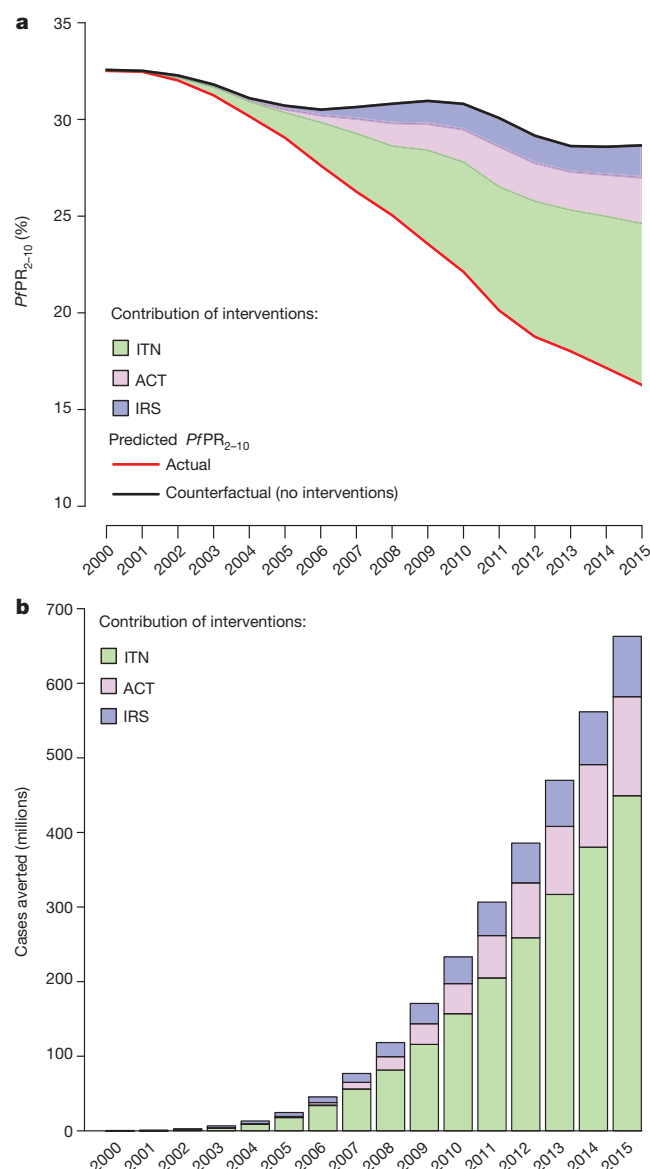


Figure 2 | Changing endemicity and effect of interventions 2000–2015.

a. Predicted time series of population-weighted mean $PfPR_{2-10}$ across endemic Africa. The red line shows the actual prediction and the black line a 'counterfactual' prediction in a scenario without coverage by ITNs, ACTs or IRS. The coloured regions indicate the relative contribution of each intervention in reducing $PfPR_{2-10}$ throughout the period. **b.** The predicted cumulative number of clinical cases averted by interventions at the end of each year, with the specific contribution of each intervention distinguished. Results shown in both panels are derived from a Bayesian geostatistical model fitted to $n = 27,573$ $PfPR$ survey points; $n = 24,868$ ITN survey points; $n = 96$ national survey reports of ACT coverage; $n = 688$ country-year reports on ITN, ACT and IRS distribution by national programs; and $n = 20$ environmental and socioeconomic covariate grids. Panel **b** additionally incorporates data from $n = 30$ active-case detection studies reporting *P. falciparum* clinical incidence.

This analysis has focused on evaluating changes in infection prevalence and clinical incidence, and we have not addressed effects on malaria mortality. Data on malaria deaths are sparse both spatially and temporally, and concerted efforts must be made to both increase data collection and improve the sensitivity and specificity of malaria death attribution. Integrating the results of the current study with existing malaria mortality estimation processes^{1,25,26}, to yield improved understanding of lives saved by malaria control, is an immediate priority.

The modelling framework presented here has been necessitated in part by the absence of detailed and robust surveillance data collected

routinely by health systems across Africa. The development of more robust surveillance systems to deliver geographically detailed and timely data on malaria incidence will be an increasingly important strand of malaria control efforts, particularly if prevalence continues to decline and identification and rapid response to individual malaria cases becomes critical to achieve elimination.

The efforts of the international community over the past 15 years have reduced malaria risk levels for many millions of people, and large regions of Africa are now in a position to consider elimination strategies. Despite this progress, many millions of people remain at risk of malaria disease and death in Africa in 2015. This analysis demonstrates that current malaria interventions have been highly effective at reducing prevalence and incidence across the continent, and provides strong support for sustaining and increasing access to these interventions as a cornerstone of post-2015 control strategies. This will need to be coupled with a redoubling of efforts to delay the spread of drug and insecticide resistance, tools for addressing the residual transmission that persists in some regions despite high vector control coverage, and concerted local programs to systematically detect and eliminate the remaining parasites.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 6 August; accepted 1 September 2015.

Published online 16 September 2015.

- World Health Organization. *World Malaria Report 2014* (World Health Organization, 2014).
- Roll Back Malaria Partnership/World Health Organization. *Global Malaria Action Plan 1 (2000–2015)* (World Health Organization, 2008).
- World Health Organization. *Global Technical Strategy for Malaria 2016–2030* (World Health Organization, 2015).
- Roll Back Malaria Partnership/World Health Organization. *Action and Investment to Defeat Malaria 2016–2030* (World Health Organization on behalf of the Roll Back Malaria Partnership Secretariat, 2015).
- Rowe, A. K. *et al.* Caution is required when using health facility-based data to evaluate the health impact of malaria control efforts in Africa. *Malar. J.* **8**, 209 (2009).
- Chizema-Kawesha, E. *et al.* Scaling up malaria control in Zambia: progress and impact 2005–2008. *Am. J. Trop. Med. Hyg.* **83**, 480–488 (2010).
- Lim, S. S. *et al.* Net benefits: a multicountry analysis of observational data examining associations between insecticide-treated mosquito nets and health outcomes. *PLoS Med.* **8**, e1001091 (2011).
- Lengeler, C. Insecticide-treated bed nets and curtains for preventing malaria. *Cochrane Database Syst. Rev.* **2**, CD000363 (2004).
- Giardina, F. *et al.* Effects of vector-control interventions on changes in risk of malaria parasitaemia in sub-Saharan Africa: a spatial and temporal analysis. *Lancet Glob. Health* **2**, e601–e615 (2014).
- Hay, S. I. *et al.* A world malaria map: *Plasmodium falciparum* endemicity in 2007. *PLoS Med.* **6**, e1000048 (2009).
- Gething, P. W. *et al.* A new world malaria map: *Plasmodium falciparum* endemicity in 2010. *Malar. J.* **10**, 378 (2011).
- Noor, A. M. *et al.* The changing risk of *Plasmodium falciparum* malaria infection in Africa: 2000–10: a spatial and temporal analysis of transmission intensity. *Lancet* **383**, 1739–1747 (2014).
- Patil, A. P. *et al.* Defining the relationship between *Plasmodium falciparum* parasite rate and clinical disease: statistical models for disease burden estimation. *Malar. J.* **8**, 186 (2009).
- Griffin, J. T., Ferguson, N. M. & Ghani, A. C. Estimates of the changing age-burden of *Plasmodium falciparum* malaria disease in sub-Saharan Africa. *Nat. Commun.* **5**, 3136 (2014).
- Smith, T. *et al.* Ensemble modeling of the likely public health impact of a pre-erythrocytic malaria vaccine. *PLoS Med.* **9**, e1001157 (2012).
- Hay, S. I. *et al.* Estimating the global clinical burden of *Plasmodium falciparum* malaria in 2007. *PLoS Med.* **7**, e1000290 (2010).
- Diggle, P. & Ribeiro, P. *Model-based Geostatistics* (Springer, 2007).
- Weiss, D. J. *et al.* Re-examining environmental correlates of *Plasmodium falciparum* malaria endemicity: a data-intensive variable selection approach. *Malar. J.* **14**, 68 (2015).
- Smith, D. L., Guerra, C. A., Snow, R. W. & Hay, S. I. Standardizing estimates of the *Plasmodium falciparum* parasite rate. *Malar. J.* **6**, 131 (2007).
- Cameron, E. *et al.* Defining the relationship between infection prevalence and clinical incidence of *Plasmodium falciparum* malaria. *Nat. Commun.* **6**, 8170 (2015).
- Wenger, E. A. & Eckhoff, P. A. A mathematical model of the impact of present and future malaria vaccines. *Malar. J.* **12**, 126 (2013).
- Battle, K. E. *et al.* Global database of *Plasmodium falciparum* and *P. vivax* incidence records from 1985–2013. *Sci. Data* **2**, 150012 (2015).
- WorldPop. Gridded population distributions. <http://www.worldpop.org.uk> (2015).

24. Cibulskis, R. E., Aregawi, M., Williams, R., Otten, M. & Dye, C. Worldwide incidence of malaria in 2009: estimates, time trends, and a critique of methods. *PLoS Med.* **8**, e1001142 (2011).
25. Liu, L. *et al.* Global, regional, and national causes of child mortality in 2000–13, with projections to inform post-2015 priorities: an updated systematic analysis. *Lancet* **385**, 430–440 (2015).
26. GBD 2013 Mortality and Causes of Death Collaborators. Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet* **385**, 117–171 (2015).

Supplementary Information is available in the online version of the paper.

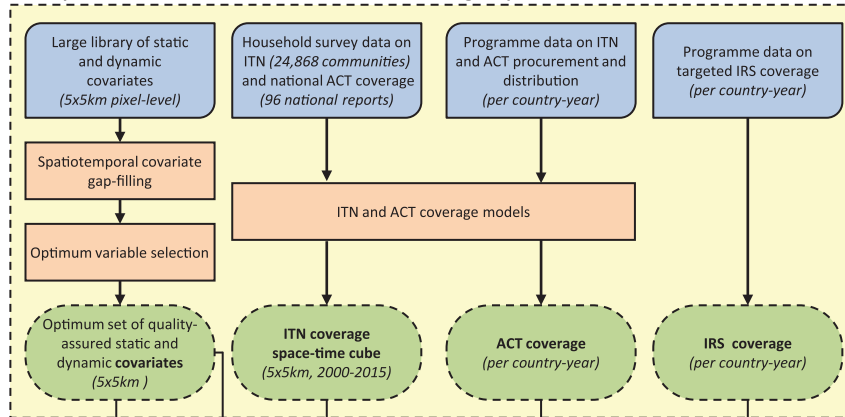
Acknowledgements The authors acknowledge assistance from M. Renshaw in providing information from the Roll Back Malaria Harmonization Working Group Programmatic Gap Analysis and other guidance in the interpretation of our results. We thank members of the Roll Back Malaria Monitoring and Evaluation Reference Group and the World Health Organization Surveillance Monitoring and Evaluation Technical expert Group for their feedback and suggestions. We thank C. Burgert of the DHS (Demographic and Health Surveys) Program for her assistance with DHS Survey access and interpretation. P.W.G. is a Career Development Fellow (no. K00669X) jointly funded by the UK Medical Research Council (MRC) and the UK Department for International Development (DFID) under the MRC/DFID Concordat agreement and receives support from the Bill and Melinda Gates Foundation (BMGF; nos OPP1068048, OPP1106023). These grants also support E.C., S.B., B.M., U.D., D.J.W.,

D.B. and A.H. The Swiss TPH component was supported through the project no. OPP1032350 funded by the BMGF. D.L.S. is funded by the BMGF (OPP1110495). S.I.H. is funded by a Senior Research Fellowship from the Wellcome Trust (no. 095066), which also supports K.E.B., and grants from the BMGF (nos. OPP1119467, OPP1106023 and OPP1093011). S.I.H. and D.L.S. also acknowledge funding support from the RAPIIDD program of the Science & Technology Directorate, Department of Homeland Security, and the Fogarty International Center, National Institutes of Health. J.T.G. is funded by an MRC Fellowship (no. G1002284). E.A.W. and P.A.E. are funded by the Global Good Fund.

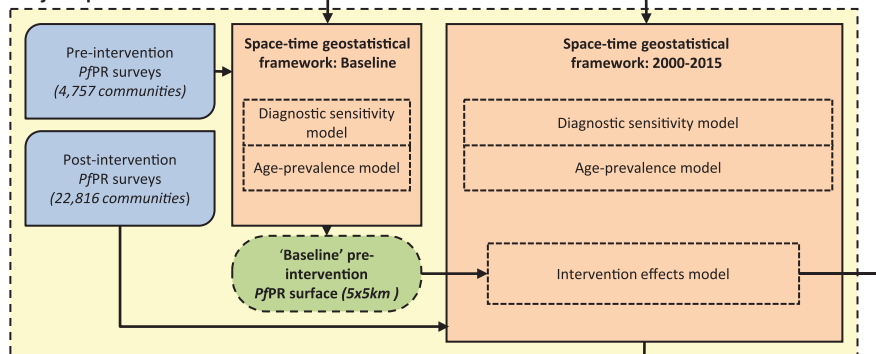
Author Contributions Conceived of and designed the research: P.W.G. and S.B. Drafted the manuscript: P.W.G. and S.B. Drafted the Supplementary Information: S.B., D.J.W., E.C., D.B., U.D., B.M. Prepared data: S.B., D.J.W., B.M., U.D., K.B., C.L.M., A.H., A.B., J.Y., T.P.E. Conducted the analyses: S.B., D.J.W., E.C., D.B., C.A.F., M.L., R.E.C. Supported the analyses: P.A.E., E.A.W., O.B., M.A.P., T.A.S., J.T.G., C.A.F., M.L., F.L., D.L.S. Supported interpretation and policy contextualization: S.B., A.B., T.P.E., J.Y., C.A.F., M.L., J.M.C., C.L.J.M., D.L.S., S.I.H., R.E.C., P.W.G. All authors discussed the results and contributed to the revision of the final manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to P.W.G. (peter.getting@zoo.ox.ac.uk).

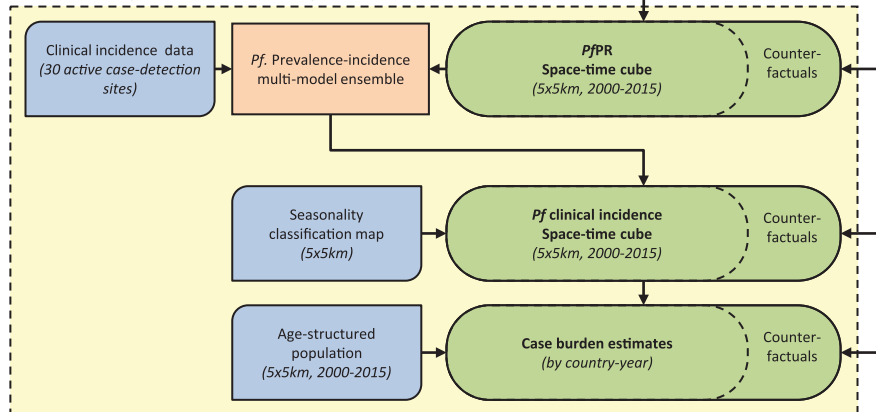
1. Preparation of covariate and intervention coverage inputs



2. PfPR space-time model



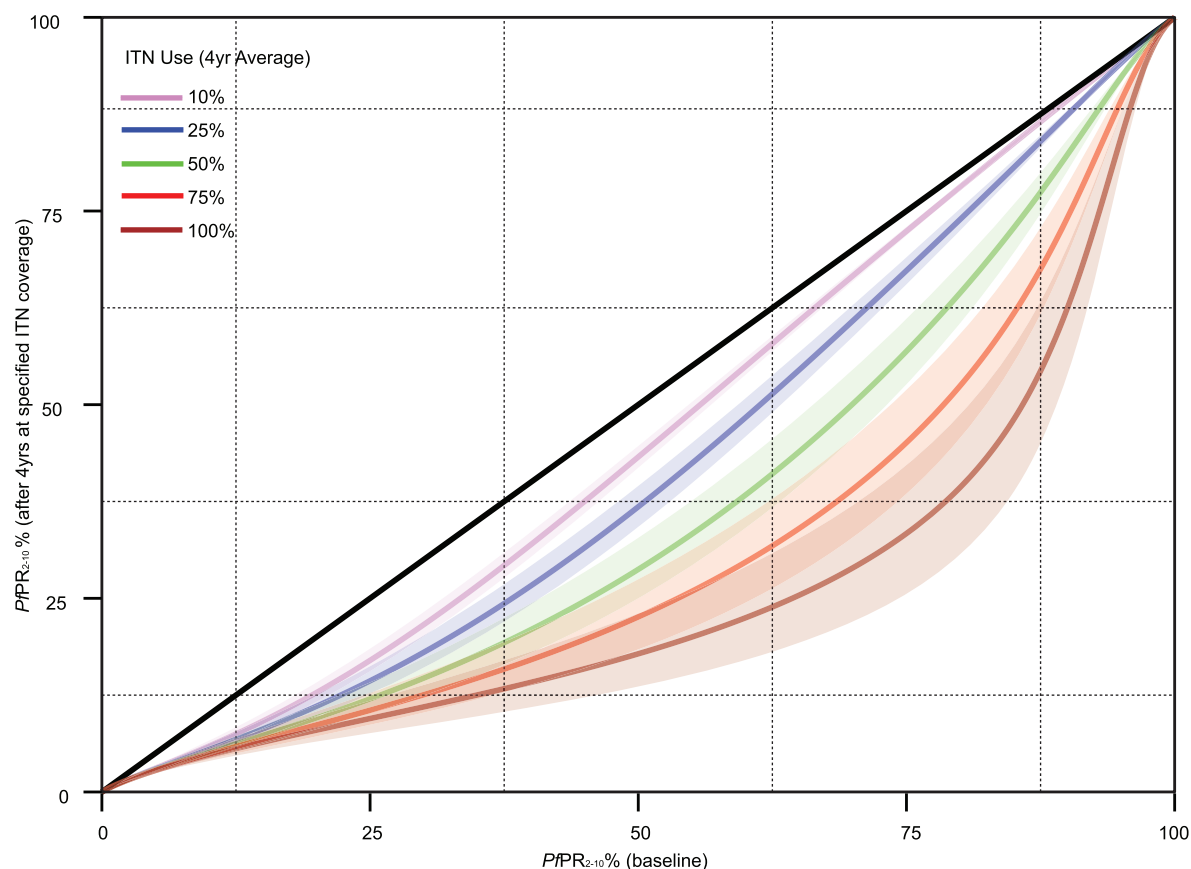
3. Predicted space-time cubes of PfPR and clinical incidence



Key



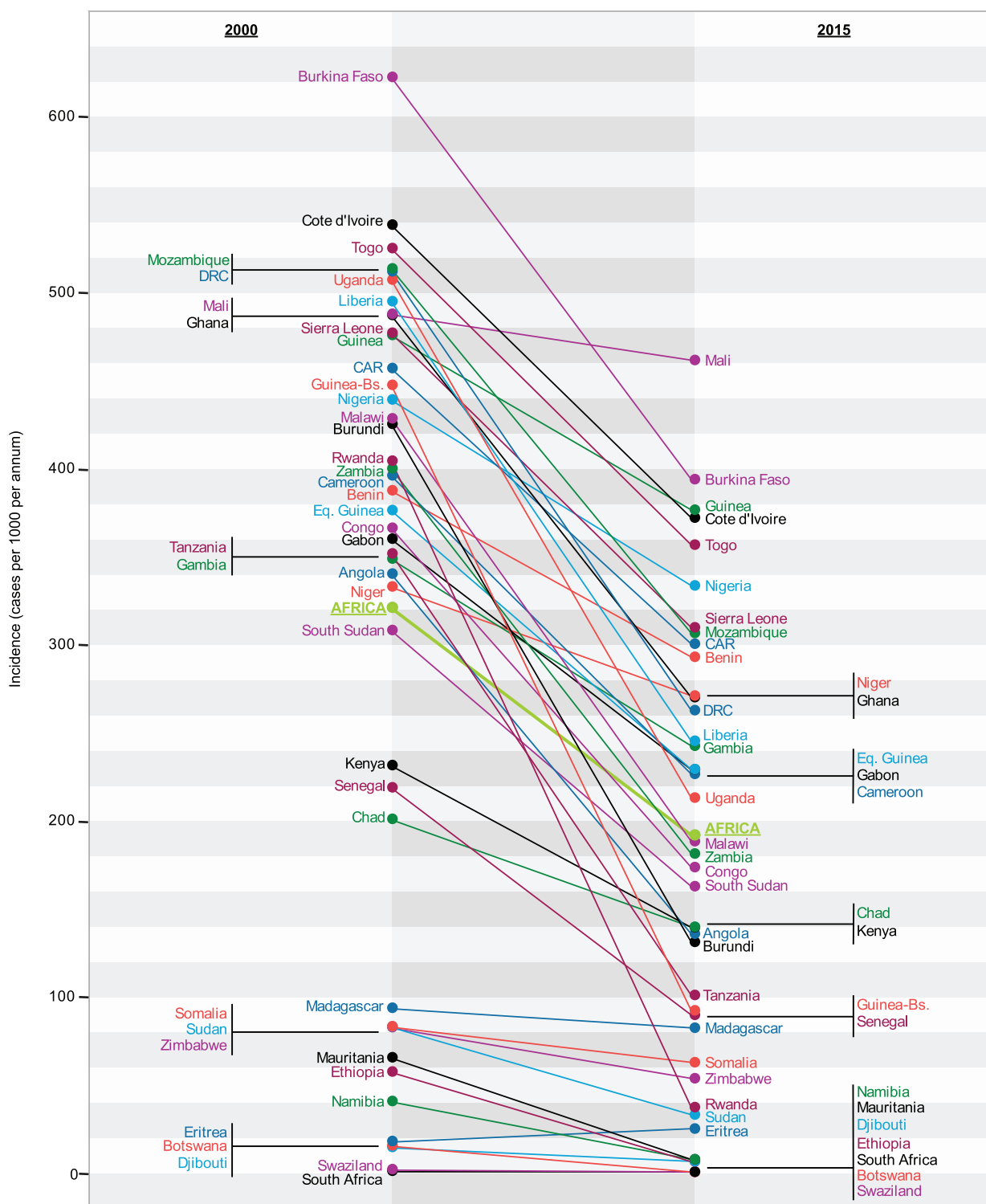
Extended Data Figure 1 | Schematic overview of main input data, model components, and outputs. Each component is detailed in the Supplementary Information.



Extended Data Figure 2 | Fitted function representing effect of ITNs.

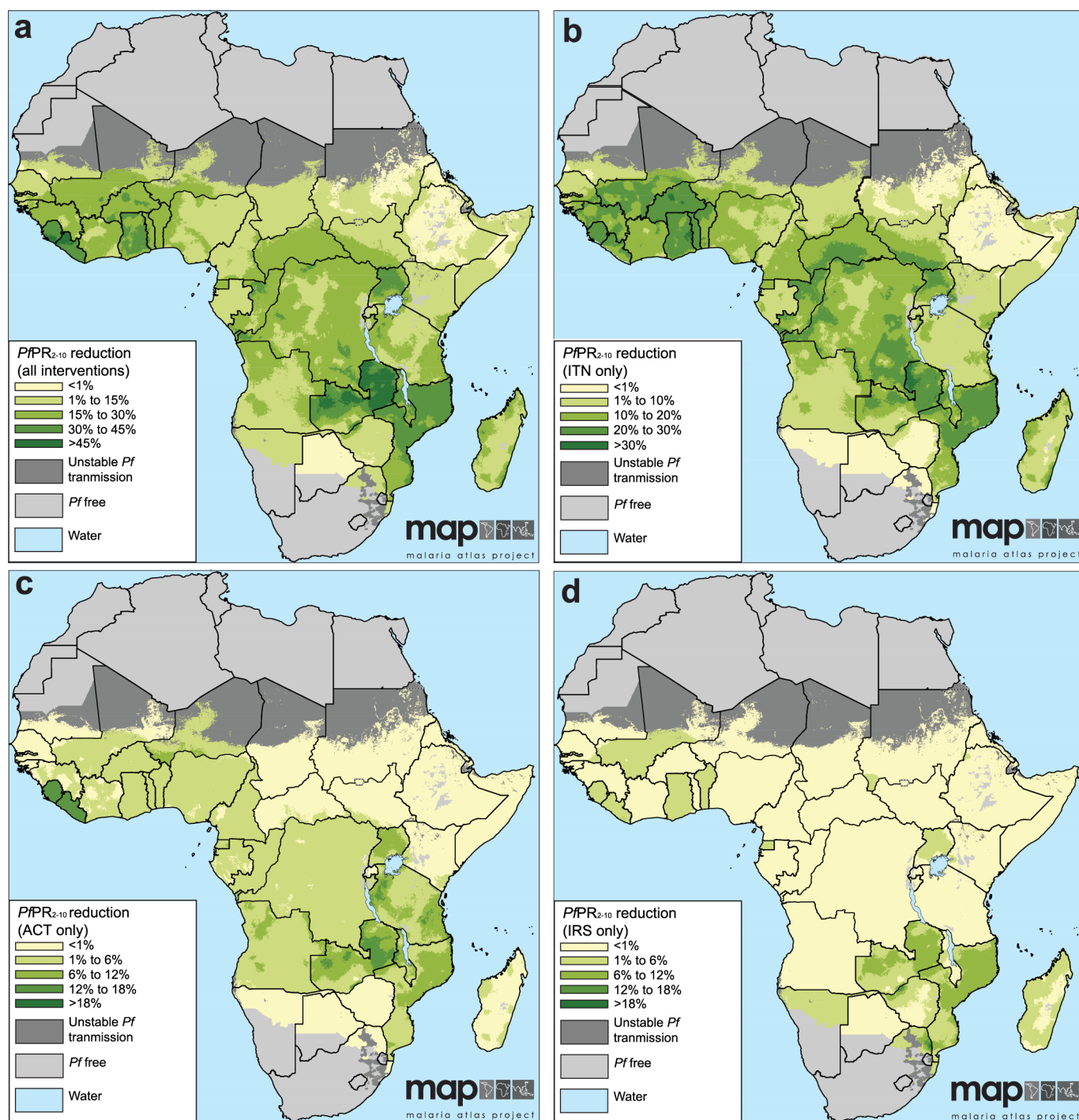
Curves illustrate the predicted effect of ITNs as a function of coverage (five example coverage levels are shown, specified as mean coverage over preceding 4-year period) and baseline transmission. The baseline *PfPR* is shown on the horizontal axis and the suppressed *PfPR* given the ITN coverage level shown on the vertical axis. The diagonal line (representing zero ITN effect) is

shown in black, and parameter uncertainty around each ITN effect line is illustrated by the semi-transparent envelopes. Results shown are derived from a Bayesian geostatistical model fitted to $n = 27,573$ *PfPR* survey points; $n = 24,868$ ITN survey points; $n = 96$ national survey reports of ACT coverage; $n = 688$ country-year reports on ITN, ACT and IRS distribution by national programs; and $n = 20$ environmental and socioeconomic covariate grids.



Extended Data Figure 3 | Changing incidence rate by country, 2000–2015. Estimated country-level rates of all-age clinical incidence are shown for 2000 and 2015. For Sudan and South Sudan, we used the post-2011 borders throughout the time period to allow comparability. Results shown are derived from a Bayesian geostatistical model fitted to $n = 27,573$ PfPR survey

points; $n = 24,868$ ITN survey points; $n = 96$ national survey reports of ACT coverage; $n = 688$ country-year reports on ITN, ACT and IRS distribution by national programs; $n = 20$ environmental and socioeconomic covariate grids; and $n = 30$ active-case detection studies reporting *P. falciparum* clinical incidence.



Extended Data Figure 4 | Decline in infection prevalence attributable to main malaria control interventions. a–d, Each map shows absolute decline in *PfPR*₂₋₁₀ between 2000 and 2015 within areas of stable transmission attributable to the combined effect of ITNs, ACTs, and IRS (a); and the individual effect of ITNs (b); ACTs (c); and IRS (d). Note that the colour scaling differs between the panels. Results shown in all panels are derived from a Bayesian geostatistical model fitted to $n = 27,573$ *PfPR* survey points;

$n = 24,868$ ITN survey points; $n = 96$ national survey reports of ACT coverage; $n = 688$ country-year reports on ITN, ACT and IRS distribution by national programs; and $n = 20$ environmental and socioeconomic covariate grids. Maps in this figure are available from the Malaria Atlas Project (<http://www.map.ox.ac.uk/>) under the Creative Commons Attribution 3.0 Unported License.

HIV-1 Nef promotes infection by excluding SERINC5 from virion incorporation

Annachiara Rosa^{1*}, Ajit Chande^{1*}, Serena Ziglio^{1*}, Veronica De Sanctis², Roberto Bertorelli², Shih Lin Goh³, Sean M. McCauley³, Anetta Nowosielska³, Stylianos E. Antonarakis^{4,5}, Jeremy Luban³, Federico Andrea Santoni⁴ & Massimo Pizzato¹

HIV-1 Nef, a protein important for the development of AIDS, has well-characterized effects on host membrane trafficking and receptor downregulation. By an unidentified mechanism, Nef increases the intrinsic infectivity of HIV-1 virions in a host-cell-dependent manner. Here we identify the host transmembrane protein SERINC5, and to a lesser extent SERINC3, as a potent inhibitor of HIV-1 particle infectivity that is counteracted by Nef. SERINC5 localizes to the plasma membrane, where it is efficiently incorporated into budding HIV-1 virions and impairs subsequent virion penetration of susceptible target cells. Nef redirects SERINC5 to a Rab7-positive endosomal compartment and thereby excludes it from HIV-1 particles. The ability to counteract SERINC5 was conserved in Nef encoded by diverse primate immunodeficiency viruses, as well as in the structurally unrelated glycosylated Gag from murine leukaemia virus. These examples of functional conservation and convergent evolution emphasize the fundamental importance of SERINC5 as a potent anti-retroviral factor.

Nef is a 27–32-kilodalton (kDa) protein expressed uniquely by primate lentiviruses that has a fundamental role in virus replication and the development of AIDS^{1–3}. It is a multifunctional factor that performs a plethora of activities within the cell, among which is the ability to downregulate crucial cell surface molecules (including CD4, MHC-I and T-cell receptor) via interaction with vesicular trafficking machinery⁴. Other activities of Nef include the ability to alter the activation state of T cells and macrophages^{5–8} and to perturb the actin cytoskeleton⁹ by engaging with cellular kinases. These relatively well-characterized activities, however, do not explain another function of Nef that was reported 20 years ago¹⁰, that is, its ability to enhance the infectivity of the virion. The latter activity seems to be important for HIV-1 pathogenesis because it is phylogenetically conserved among widely divergent primate lentiviruses¹¹ and maintained under strong selective pressure during disease progression¹². Such enhancement of virion infectivity depends on *nef* being expressed from within virus-producing cells¹³, but it is manifest at an early stage in the subsequent infection of susceptible target cells^{13–15}, indicating a yet unknown modification of progeny virus particles.

Although Nef is unique to HIV and SIV, glycosylated Gag from an unrelated gammaretrovirus (Moloney murine leukaemia (MLV)) fully substitutes for the activity of Nef on HIV-1 infectivity¹⁶. Despite the lack of any sequence homology, Nef and glycosylated Gag share a remarkable functional similarity, as they both require host cell endocytosis machinery to boost virion infectivity¹⁷. A Nef-like activity promoting retrovirus infectivity has therefore arisen by convergent evolution within an unrelated family of retroviruses. However, the molecular mechanism underlying the requirement of Nef and glycosylated Gag for retrovirus infectivity has so far remained elusive.

Nef counteracts a retrovirus inhibitor

We investigated to what extent the Nef requirement for virion infectivity is producer cell-type dependent, by comparing the infectivity of wild-type HIV-1 to its Nef-defective counterpart produced from 31 different

human cell lines (Fig. 1a and Extended Data Table 1). Varying with the producer cell type, the effect of Nef ranged from 2- to 40-fold, arguing in favour of the presence of a cellular inhibitor of HIV-1 counteracted by Nef. We then investigated whether this Nef responsiveness is a dominant feature in producer cells by generating Nef-positive and Nef-negative HIV-1 virions from heterokaryons derived from cell lines with opposite Nef-responsiveness (Fig. 1b). When lymphoid cells (high Nef responsive) were fused with fibrosarcoma cells (low Nef responsive), HIV-1 produced by heterokaryons displayed relatively high dependence on Nef (Fig. 1c), indicating the presence of a transdominant cellular inhibitor of HIV-1 infectivity counteracted by Nef.

To identify such a putative host factor, the global transcriptome of high and low Nef-responsive cells was examined to pinpoint differentially expressed genes that correlate with Nef responsiveness. Transcriptomes from seven highly Nef-responsive cell lines (Nef effect ranging from 10- to 40-fold) and eight low Nef-responsive cell lines (Nef effect lower than fourfold) were subjected to RNA-sequencing (RNA-seq). On the basis of correlation analysis, SERINC5 emerged as the gene whose expression correlated best with the requirement of Nef for HIV-1 infectivity (Fig. 1d).

SERINC5 inhibits HIV-1 and MLV

To validate functionally the effect on virion infectivity, the *SERINC5* genomic sequence was disrupted in the cell line with the highest Nef responsiveness (Jurkat TAg or JTAG) using a clustered regularly interspaced short palindromic repeat (CRISPR)-Cas9 lentiviral vector (Extended Data Fig. 1a). SERINC5 knockout cells produced a 20–30-fold increase in the infectivity of the Nef-defective HIV-1, whereas the Nef-positive virus was only affected 2–3-fold, thus reducing the Nef effect from 50- to 3-fold (Fig. 2a, b). This result was reproduced targeting three different regions of the *SERINC5* gene (Extended Data Fig. 1b). When haemagglutinin (HA)-tagged SERINC5 was expressed from a complementary DNA non-targetable by the CRISPR-Cas9 vector, the high Nef-dependent phenotype was

¹University of Trento, Centre for Integrative Biology, 38123 Trento, Italy. ²University of Trento, Laboratory of Biomolecular Sequence and Structure Analysis for Health, NGS facility, 38123 Trento, Italy. ³University of Massachusetts Medical School, Program in Molecular Medicine, Worcester, Massachusetts 01605, USA. ⁴University of Geneva, Department of Genetic Medicine and Development, Geneva 1211, Switzerland. ⁵GE3 Institute of Genetics and Genomics of Geneva, Geneva 1211, Switzerland.

*These authors contributed equally to this work.

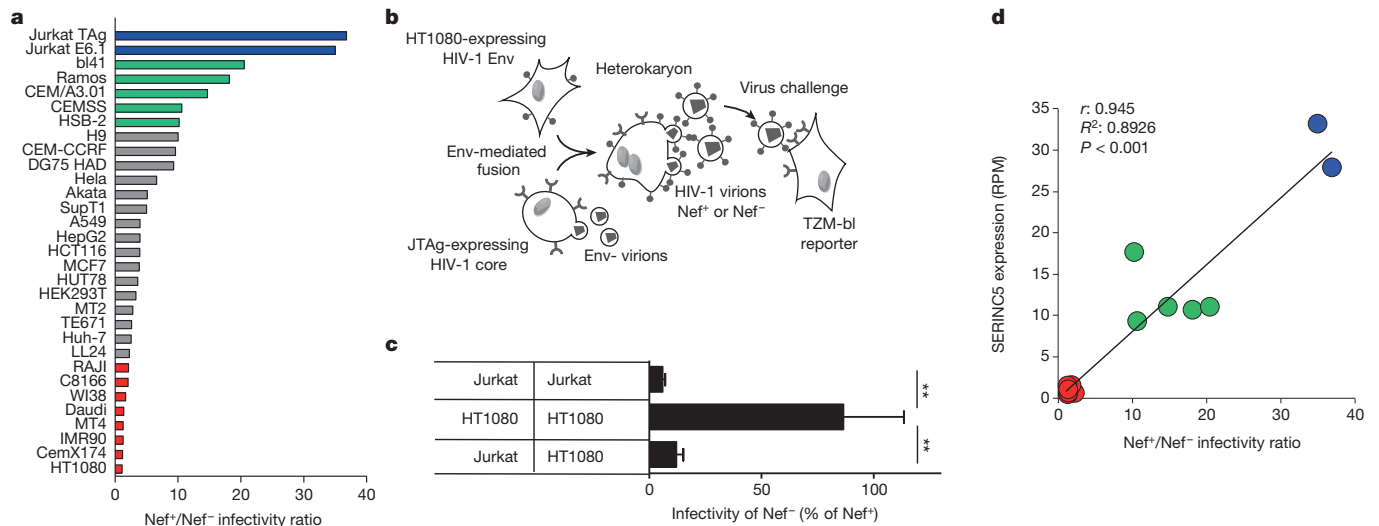


Figure 1 | Nef counteracts an HIV-1 inhibitor. **a**, Ratio of the infectivity of NL4-3 and NL4-3^{Nef-} produced from the indicated cell lines and measured on TZM-bl reporter cells. **b**, The schematic of the heterokaryon assay. **c**, Infectivity of HIV-1 derived from heterokaryons generated by the indicated cell lines ($n = 3$, mean \pm s.d., unpaired t -test, $**P < 0.01$).

d, Correlation of SERINC5 expression in producer cells and Nef requirement for infectivity. Colours in **a** and **d** represent the same cell lines. Trendline indicates linear regression. (Pearson correlation, two-tailed, $P < 0.0001$). RPM, reads per million.

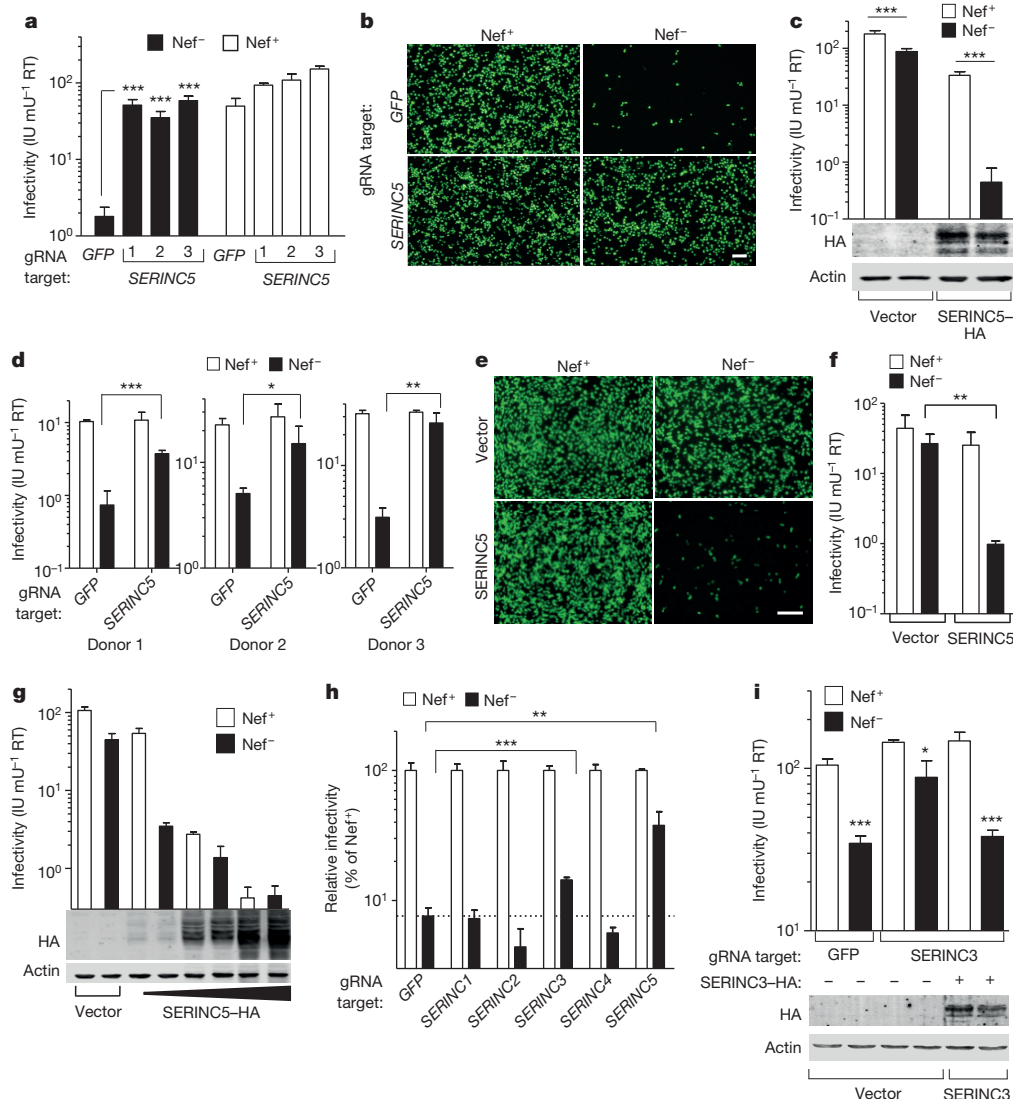


Figure 2 | SERINC5 and SERINC3 inhibit HIV-1. **a**, Infectivity of HIV-1 from cells stably transduced with lentiCRISPR ($n = 4$). gRNA, guide RNA; RT, reverse transcriptase. **b**, Fluorescence microscopy of reporter cells from **a**. **c**, Infectivity of HIV-1 from JTag SERINC5^{-/-} and immunoblotting of producer cells. **d**, Infectivity of HIV-1 from PBMC, co-transfected with CRISPR-Cas9 vectors ($n = 3$, one experiment performed per donor). **e-g**, Infectivity of HIV-1 from HEK293T expressing SERINC5-HA. Reporter cells infected with HIV-1 from HEK293T expressing SERINC5-HA ($n = 4$). **h, i**, Infectivity of HIV-1 produced from JTag (h) and JTag SERINC5^{-/-} (i) transfected with CRISPR-Cas9 vectors targeting the indicated SERINC genes ($n = 4$). In **i**, immunodetection of SERINC3-HA in producer cells. Mean \pm s.d., unpaired two-tailed t -test, $*P < 0.05$, $**P < 0.01$, $***P < 0.001$. Scale bars, 100 μ m.

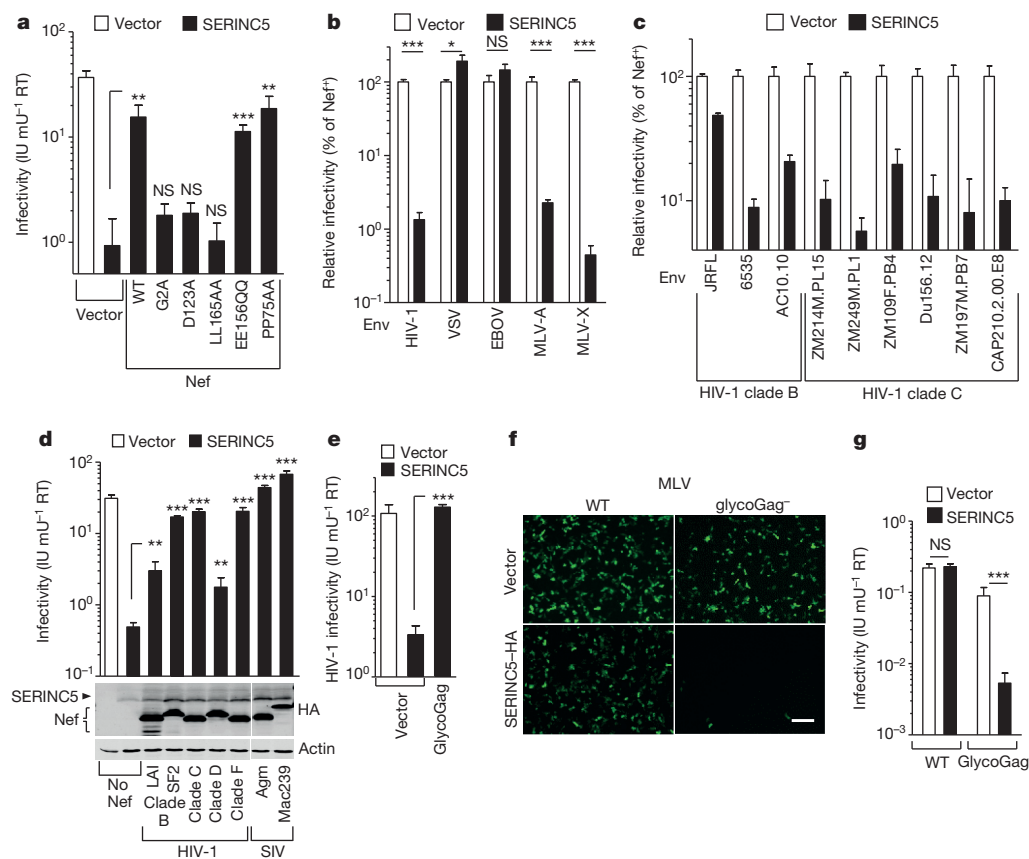


Figure 3 | Determinants of Nef activity against SERINC5 and conservation across different retroviruses. **a**, The ability of Nef mutants to counteract SERINC5 inhibition of HIV-1 infectivity ($n = 4$). **b**, **c**, Susceptibility of viral pseudotypes to inhibition of infectivity by SERINC5 ($n = 4$). **d**, **e**, Counteraction of SERINC5 by *nef* alleles and immunoblot from producer cells (**d**) and glycoGag (**e**) on HIV-1. **f**, **g**, Infectivity of wild-type and glycoGag-defective MLV from HEK293T expressing SERINC5 ($n = 4$). Mean \pm s.d., unpaired two-tailed *t*-test, * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$; NS, not significant. Scale bar, 100 μ m.

restored (Fig. 2c), and the infectivity of the Nef-defective HIV-1 was reduced 197-fold versus a fivefold only reduction of the Nef-positive counterpart. SERINC5 was found to be expressed in primary blood cells from three different donors to a level comparable to that observed in Jurkat cells (Extended Data Fig. 1c). Accordingly, CRISPR-Cas9 vector-mediated SERINC5 knockout cells increased specifically the infectivity of Nef-defective HIV-1 produced in cultured peripheral blood mononuclear cells (PBMC) derived from three different individuals (Fig. 2d), demonstrating that SERINC5 inhibits HIV-1 produced in primary human blood cells.

Ectopic expression of SERINC5 in cells with minimal Nef-dependence (Fig. 2e, f and Extended Data Fig. 1d), resulted in a 10–40-fold selective inhibition of Nef-defective HIV-1. SERINC5 is therefore not only required, but also sufficient to inhibit HIV-1 infectivity and to confer Nef responsiveness. While inhibition of HIV-1 infection by SERINC5 is dose-dependent (Fig. 2g), the ability of Nef to preserve the infectivity of the virus particle is abolished with increasing expression of SERINC5, suggesting that the ability of Nef to counteract SERINC5 is saturable (Fig. 2g). At the highest SERINC5 expression level, virion infectivity was reduced 256-fold, regardless of Nef expression.

SERINC5 belongs to a unique gene family present in all eukaryotes and contains 10 putative transmembrane helices^{18,19}. While it was suggested that SERINC proteins mediate incorporation of serine into membrane lipids²⁰, their function is unknown. The five members of the human SERINC family share more than 17% amino acid identity and a similarly predicted membrane topology. We observed that virus produced in JTag *SERINC5*^{-/-} cells retains a 2–3-fold responsiveness to Nef (Fig. 2a). Our transcriptome analysis indicated that JTag cells express other *SERINC* genes in addition to *SERINC5* (Extended Data Fig. 1e). We therefore explored the possibility that other SERINC family members have anti-HIV-1 activity by knocking out the five *SERINC* genes individually. Targeting *SERINC3* in JTag *SERINC5*^{-/-} cells resulted in a 2–3-fold rescue of Nef-defective virus

infectivity (Fig. 2i), thus further reducing the residual Nef responsiveness to 1.6-fold (Fig. 2i). Ectopic expression of SERINC3 resulted in threefold inhibition of Nef-defective HIV-1 (Fig. 2i), confirming that SERINC3 can also inhibit HIV-1 infectivity.

The Nef activity against SERINC5

The effect of Nef on infectivity requires Nef myristoylation and interaction with clathrin-mediated endocytosis^{21,22} (AP2 and dynamin2). Accordingly, the ability to counteract SERINC5 was impaired by *nef* mutations that abolish Nef amino-terminal myristoylation (G2A), disrupt a di-leucine-based sorting signal (LL165AA) necessary for AP2 interaction²¹, or prevent binding to dynamin 2 (D123A, Fig. 3a)²². By contrast, mutations abrogating either a proline-rich SH3 binding domain (PP75AA)²³, or di-acidic motif required for CD4 downregulation (EE156QQ)²⁴, do not affect the ability to counteract SERINC5 (Fig. 3a). The molecular features of Nef already known to be crucial for the effect on infectivity are therefore required for counteracting SERINC5.

It has been reported that the effect of Nef on infectivity depends on the nature of the envelope glycoprotein^{16,25–29}. Accordingly, pseudotyping HIV-1 with vesicular stomatitis virus G protein (VSV-G) and with the Ebola virus glycoprotein (EBOV GP), but not with MLV-A nor MLV-X Env, makes HIV-1 resistant to SERINC5 (Fig. 3b). The magnitude of the effect of Nef on infectivity was also reported to vary when HIV-1 is pseudotyped with envelope glycoproteins derived from different HIV-1 isolates^{30,31}. Accordingly, virions carrying Env derived from a panel of HIV-1 primary isolates were variably affected by SERINC5 (Fig. 3c), indicating that naturally occurring isolates are inhibited by the host factor to different extents.

The activity of Nef on infectivity is highly conserved among primate lentiviruses¹¹. We therefore tested whether the ability to counteract SERINC5 is shared among *nef* alleles. Nef proteins derived from subtypes B, C, D and F clinical isolates could counteract ectopically

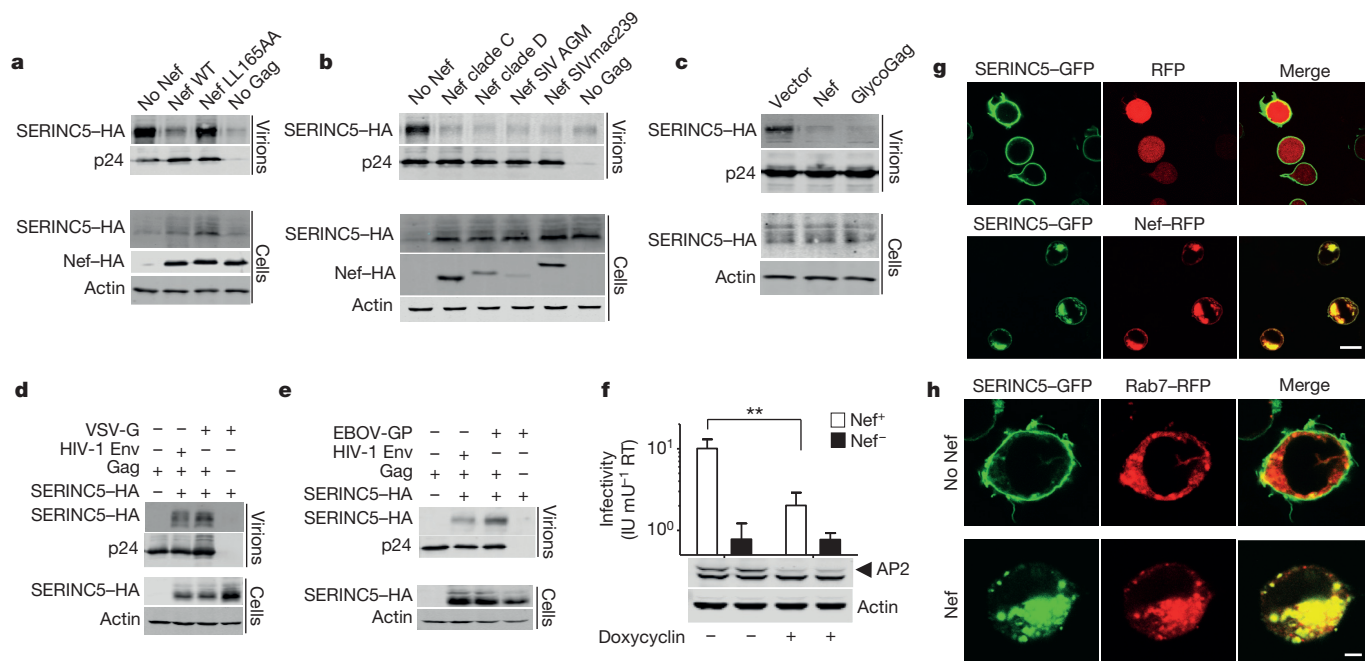


Figure 4 | Nef and glycoGag promote relocalization of SERINC5 to an endosomal compartment and prevent its incorporation into virions. **a–e**, Immunoblots on viral particles and corresponding cell lysates from *nef*-defective NL4-3 complemented with plasmids encoding Nef proteins as indicated (**a** and **b**), MLV glycoGag (**c**), VSV-G (**d**) EBOV-GP (**e**) and a vector expressing SERINC5–HA. **f**, Infectivity of HIV-1 from HEK293T cells stably

expressing a doxycycline-inducible shRNA targeting AP2 and transfected with PBJ6–SERINC5. Western blot: AP2 in cell lysates derived from producer cells (mean \pm s.d., $n = 4$ unpaired two-tailed t -test, $**P < 0.01$, experiment replicated twice). **g**, **h**, Confocal microscopy of JTAG cells transfected to express SERINC5–GFP with RFP, Nef–RFP (**g**) or Rab7–RFP (**h**). Scale bars, 10 μ m (**g**) and 2 μ m (**h**).

expressed SERINC5–HA with a potency 5–10-fold higher than that observed with Nef derived from a laboratory adapted strain (HIV-1_{LAI}, Fig. 3d). Similarly, Nef from two divergent SIV lineages (SIVmac239 and SIVagm) also counteracted SERINC5 with tenfold higher efficacy than HIV-1_{LAI} (Fig. 3d). The ability to counteract SERINC5 is therefore a prominent feature of Nef, conserved across different primate lentivirus species.

We next tested whether SERINC5 can target retroviruses other than lentiviruses. We have shown that glycosylated Gag (glycoGag) from MLV is capable of rescuing the infectivity of Nef-defective HIV-1 (ref. 16), despite sharing no sequence homology with Nef. Indeed, glycoGag efficiently rescues the infectivity of HIV-1 (Fig. 3e) by counteracting SERINC5, suggesting that SERINC5 has an important role also in the context of infection with gammaretroviruses. Accordingly, SERINC5 expression in producer cells potently inhibited infectivity of MLV only in the absence of glycoGag (Fig. 3f, g). Therefore, while SERINC5 targets divergent retroviruses, factors capable of overcoming its inhibitory activity on infectivity have evolved independently.

Incorporation of SERINC5 into virions

The ability of SERINC5 to be incorporated into the lipid envelope of HIV-1 virions was tested next. HIV-1 was produced in JTAG SERINC5^{-/-} expressing SERINC5–HA. Despite being barely detectable in cells in the absence of Nef, SERINC5–HA was readily visualized in Nef-defective virions and was largely excluded from virions generated in the presence of Nef (Fig. 4a) but not in the presence of the Nef mutant lacking the AP2 binding site (LL165AA, Fig. 4a). The ability to prevent virion incorporation of SERINC5 was readily observed with Nef alleles from HIV-1 and SIV (Fig. 4b) and with MLV glycoGag (Fig. 4c), suggesting that association with virions is crucial for the effect on infectivity and is tightly controlled by both primate lentiviral and gammaretroviral factors. The effect of Nef on SERINC5 association with virions did not alter the amount of incorporated Env (Extended Data Fig. 2a), in line with previous

observations that failed to observe any effect of Nef on virion Env abundance^{16,25,31}. By contrast, the amount of SERINC5 incorporated into HIV particles was not reduced by VSV-G (Fig. 4d) nor by EBOV GP (Fig. 4e), despite the infectivity of VSV-G and EBOV GP pseudotypes being resistant to the host factor (Fig. 3b). Therefore, while Nef and glycoGag seem to counteract SERINC5 by preventing its incorporation into virions, VSV-G and EBOV GP must antagonize its effect by a different mechanism.

The ability of Nef to counteract SERINC5 was significantly reduced by silencing AP2 (Fig. 4f), confirming the crucial involvement of clathrin-dependent intravesicular trafficking. Using immunofluorescence microscopy, green fluorescent protein (GFP)-tagged SERINC5 (Fig. 4g) was observed to localize almost exclusively to the plasma membrane. By contrast, the expression of HIV-1 Nef caused SERINC5 to relocalize together with Nef into perinuclear vesicles identified as late endosomes (RAB7-positive, Fig. 4h). SERINC5 was similarly efficiently retargeted into perinuclear vesicles by SIV Nef and by MLV glycoGag (Extended Data Fig. 2b), indicating a common ability of the retroviral factors to relocalize SERINC5, which is removed from the plasma membrane, and prevented from accessing nascent virions.

The anti-HIV-1 activity of SERINC5

Which step of the HIV-1 life cycle is blocked by SERINC5 was investigated next. HIV-1 produced in the presence of SERINC5–HA failed to accumulate products of reverse transcription in target cells, confirming previous reports that in the absence of Nef the infection process is halted at an early stage of the HIV-1 life cycle (Fig. 5a). Whether Nef affects fusion between the virion particle and the target cell membrane has remained questionable^{25,32–37}. We therefore developed a novel protein transduction assay in which the bacteriophage Cre recombinase fused to a nuclear localization signal flanked by HIV-1 protease cleavage sites (Fig. 5b) is packaged as part of the Gag polypeptide into HIV-1 particles (Extended Data Fig. 3a). Cre, delivered into the cell after fusion, activates expression of a reporter

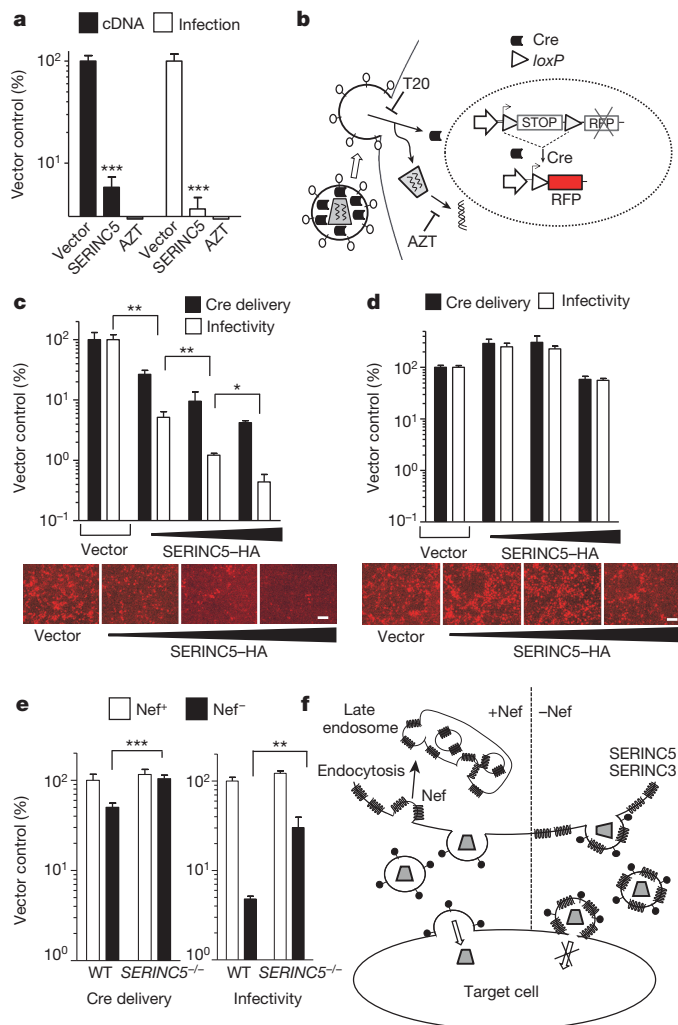


Figure 5 | SERINC5 inhibits an early step of virus infection. **a**, Effect of SERINC5 on the generation of HIV-1 late reverse transcription products ($n = 3$; experiments replicated twice) and the corresponding effect on infectivity. **b**, Schematic of the nlsCre delivery assay. **c**, **d**, Effect of SERINC5 on Cre-delivery by HIV-1 (**c**) and HIV-1 pseudotyped with VSV-G (**d**). **e**, Cre delivery and infectivity by HIV-1-derived from JTAG or JTAG SERINC5^{-/-}. **f**, Schematic showing the activity of SERINC5 on HIV-1 infectivity and the counteracting mechanism by Nef. Mean \pm s.d., $n = 4$, unpaired two-tailed t -test, * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$. Scale bars, 100 μ m.

gene (nlsRFP) following *loxP* recombination. Confirming the ability to detect cytoplasmic delivery of Cre independently of productive infection, Cre-mediated reporter activation was not blocked by the reverse transcriptase inhibitor azidothymidine (AZT) but was blocked by the fusion inhibitor T20 (Extended Data Fig. 3b).

Increasing expression of SERINC5 in producer cells did not affect the amount of Cre associated with virions (Extended Data Fig. 3c), but resulted in a gradually increased inhibition of Cre-mediated activation of the reporter gene in target cells by Nef-defective HIV-1 (Fig. 5c), with a 25-fold inhibition observed at the highest SERINC5 expression level, which in turn inhibited infectivity by 250-fold (Fig. 5c). This observation was reproducible also using a fusion assay based on the viral incorporation and cytoplasmic delivery of a BLAM-VpR chimaeric gene³⁸ (Extended Data Fig. 3d). By contrast, Cre-delivery from Nef-defective HIV-1 pseudotyped with VSV-G (Fig. 5d) or with EBOV GP (Extended Data Fig. 3e) was not inhibited by SERINC5, consistent with the

intrinsic resistance of these pseudotypes to the inhibition by the host factor (Fig. 3b).

When the host factor was expressed at a level which introduced a 20-fold effect on infectivity, SERINC5 resulted in a 2–3-fold inhibition of Cre delivery, fully counteracted by Nef (Extended Data Fig. 3f). Similarly, Nef-defective HIV-1 derived from wild-type JTAG cells delivered Cre to target cells with a 2–3-fold lower efficiency than Nef-positive virions in the presence of endogenously expressed SERINC5, in spite of a 20-fold lower infectivity (Fig. 5e).

Altogether, these results suggest that SERINC5 perturbs the ability of the viral particle to translocate its content to the cytoplasm.

Discussion

Here we demonstrate that SERINC5, and to a lesser extent SERINC3, are responsible for the long-sought anti-HIV-1 activity that is overcome by Nef. These cellular proteins join a growing list of host factors that inhibit retrovirus infection and are referred to as restriction factors. However, SERINC5 and SERINC3 have features that distinguish them from other known retroviral restriction factors. For example, SERINC5 expression in primary CD4⁺ T cells or dendritic cells is not upregulated by type I interferon or by an interferon-inducing agent such as lipopolysaccharide (LPS; Extended Data Fig. 4). SERINC5 and SERINC3 therefore appear to be examples of constitutively expressed intrinsic restriction factors.

Human SERINC5 shares 28% identity at the amino acid level with the *Saccharomyces cerevisiae* orthologue, *TMS1* (ref. 18). Such a degree of conservation suggests a yet unidentified core biological function in cells and represents another peculiar feature compared with other antiretroviral restriction factors (for example, TRIM5 could be traced back only to teleosts³⁹), which diversified under strong positive selection⁴⁰. Remarkably, Nef from HIV-1 and SIV, as well as glycoGag from MLV, are all capable of counteracting human SERINC5, denoting an unusual low species-specificity between the host factor and the viral antagonist.

We provided evidence that SERINC5 perturbs the ability of small intravirion proteins, such as Cre and BLAM-VpR (less than 40 kDa), to access the target cell. However, inhibition of infectivity by SERINC5 is tenfold higher, suggesting that infection is blocked despite detectable fusion. The effect on infectivity, which requires the delivery of the 60–120-nm viral core⁴¹, is therefore unlikely to be explained only by an effect of SERINC5 on the initial membrane fusion event. The host protein could therefore affect a step after the fusion pore generation, required for the translocation of the viral core (Fig. 5f). After the initial membrane fusion triggered by fusogenic glycoproteins, the formation of a fusion pore is followed by its expansion, the highest energy requiring step in the fusion process⁴². This event is known to be affected by the lipid membrane composition^{43,44} and the presence of proteins altering the rigidity and the curvature of the lipid bilayer⁴⁵. How SERINC5 would affect this step of HIV-1 infection remains to be established. By contrast, VSV-G or EBOV GP may override such inhibition by intrinsically promoting more efficient expansion of the fusion pore. Interestingly, some HIV-1 Env glycoproteins from clinical isolates appear also to modulate the susceptibility of the virus to SERINC5 inhibition (Fig. 3c), suggesting the possibility that HIV-1 uses Env in addition to Nef to overcome such a powerful block.

In conclusion, the ability to target evolutionary distant retroviruses (HIV and MLV) and the convergent evolution of antagonistic retroviral factors (Nef and glycoGag) indicate that SERINC5 has a fundamental role in the interaction of the host with retroviral pathogens. Interestingly, ectopic expression of SERINC5 potently inhibits HIV-1, even in the presence of Nef (Fig. 2g), suggesting that this cellular antiviral factor might be exploited as an anti-HIV-1 therapeutic gene.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 6 May; accepted 18 August 2015.

Published online 30 September 2015.

1. Kestler, H. W. Importance of the *nef* gene for maintenance of high virus loads and for development of AIDS. *Cell* **65**, 651–662 (1991).
2. Deacon, N. J. *et al.* Genomic structure of an attenuated quasi species of HIV-1 from a blood transfusion donor and recipients. *Science* **270**, 988–991 (1995).
3. Kirchhoff, F., Greenough, T. C., Brettler, D. B., Sullivan, J. L. & Desrosiers, R. C. Brief report: absence of intact *nef* sequences in a long-term survivor with nonprogressive HIV-1 infection. *N. Engl. J. Med.* **332**, 228–232 (1995).
4. Landi, A., Iannucci, V., Nuffel, A. V., Meuwissen, P. & Verhasselt, B. One protein to rule them all: modulation of cell surface receptors and molecules by HIV Nef. *Curr. HIV Res.* **9**, 496–504 (2011).
5. Baur, A. S. *et al.* HIV-1 Nef leads to inhibition or activation of T cells depending on its intracellular localization. *Immunity* **1**, 373–384 (1994).
6. Schragar, J. A. & Marsh, J. W. HIV-1 Nef increases T cell activation in a stimulus-dependent manner. *Proc. Natl Acad. Sci. USA* **96**, 8167–8172 (1999).
7. Alexander, L., Du, Z., Rosenzweig, M., Jung, J. U. & Desrosiers, R. C. A role for natural simian immunodeficiency virus and human immunodeficiency virus type 1 *nef* alleles in lymphocyte activation. *J. Virol.* **71**, 6094–6099 (1997).
8. Simmons, A., Aluvihare, V. & McMichael, A. Nef triggers a transcriptional program in T cells imitating single-signal T cell activation and inducing HIV virulence mediators. *Immunity* **14**, 763–777 (2001).
9. Stolp, B. *et al.* HIV-1 Nef interferes with host cell motility by deregulation of Cofilin. *Cell Host Microbe* **6**, 174–186 (2009).
10. Chowes, M. Y. *et al.* Optimal infectivity *in vitro* of human immunodeficiency virus type 1 requires an intact *nef* gene. *J. Virol.* **68**, 2906–2914 (1994).
11. Munch, J. *et al.* Nef-mediated enhancement of virion infectivity and stimulation of viral replication are fundamental properties of primate lentiviruses. *J. Virol.* **81**, 13852–13864 (2007).
12. Carl, S. *et al.* Modulation of different human immunodeficiency virus type 1 Nef functions during progression to AIDS. *J. Virol.* **75**, 3657–3665 (2001).
13. Aiken, C. & Trono, D. Nef stimulates human immunodeficiency virus type 1 proviral DNA synthesis. *J. Virol.* **69**, 5048–5056 (1995).
14. Chowes, M. Y., Pandori, M. W., Spina, C. A., Richman, D. D. & Guatelli, J. C. The growth advantage conferred by HIV-1 *nef* is determined at the level of viral DNA formation and is independent of CD4 downregulation. *Virology* **212**, 451–457 (1995).
15. Schwartz, O., Marechal, V., Danos, O. & Heard, J. M. Human immunodeficiency virus type 1 Nef increases the efficiency of reverse transcription in the infected cell. *J. Virol.* **69**, 4053–4059 (1995).
16. Pizzato, M. MLV glycosylated-Gag is an infectivity factor that rescues Nef-deficient HIV-1. *Proc. Natl Acad. Sci. USA* **107**, 9364–9369 (2010).
17. Usami, Y., Popov, S. & Gottlinger, H. G. The Nef-Like effect of murine leukemia virus glycosylated Gag on HIV-1 infectivity is mediated by its cytoplasmic domain and depends on the AP-2 adaptor complex. *J. Virol.* **88**, 3443–3454 (2014).
18. Grossman, T. R., Luque, J. M. & Nelson, N. Identification of a ubiquitous family of membrane proteins and their expression in mouse brain. *J. Exp. Biol.* **203**, 447–457 (2000).
19. Xu, J. *et al.* Cloning and expression of a novel human *C5orf12* gene, a member of the TMS-TDE family. *Mol. Biol. Rep.* **30**, 47–52 (2003).
20. Inuzuka, M., Hayakawa, M. & Ingi, T. Serinc, an activity-regulated protein family, incorporates serine into membrane lipid synthesis. *J. Biol. Chem.* **280**, 35776–35783 (2005).
21. Craig, H. M., Pandori, M. W. & Guatelli, J. C. Interaction of HIV-1 Nef with the cellular dileucine-based sorting pathway is required for CD4 down-regulation and optimal viral infectivity. *Proc. Natl Acad. Sci. USA* **95**, 11229–11234 (1998).
22. Pizzato, M. *et al.* Dynamin 2 is required for the enhancement of HIV-1 infectivity by Nef. *Proc. Natl Acad. Sci. USA* **104**, 6812–6817 (2007).
23. Saksela, K., Cheng, G. & Baltimore, D. Proline-rich (PxxP) motifs in HIV-1 Nef bind to SH3 domains of a subset of Src kinases and are required for the enhanced growth of Nef⁺ viruses but not for down-regulation of CD4. *EMBO J.* **14**, 484–491 (1995).
24. Piguet, V. *et al.* Nef-induced CD4 degradation: a diacidic-based motif in Nef functions as a lysosomal targeting signal through the binding of β -COP in endosomes. *Cell* **97**, 63–73 (1999).
25. Miller, M. D., Warmerdam, M. T., Page, K. A., Feinberg, M. B. & Greene, W. C. Expression of the human immunodeficiency virus type 1 (HIV-1) *nef* gene during HIV-1 production increases progeny particle infectivity independently of gp160 or viral entry. *J. Virol.* **69**, 579–584 (1995).
26. Aiken, C. Pseudotyping human immunodeficiency virus type 1 (HIV-1) by the glycoprotein of vesicular stomatitis virus targets HIV-1 entry to an endocytic pathway and suppresses both the requirement for Nef and the sensitivity to cyclosporin A. *J. Virol.* **71**, 5871–5877 (1997).
27. Chazal, N., Singer, G., Aiken, C., Hammarskjöld, M. L. & Rekosh, D. Human immunodeficiency virus type 1 particles pseudotyped with envelope proteins that fuse at low pH no longer require Nef for optimal infectivity. *J. Virol.* **75**, 4014–4018 (2001).
28. Pizzato, M., Popova, E. & Gottlinger, H. G. Nef can enhance the infectivity of receptor-pseudotyped human immunodeficiency virus type 1 particles. *J. Virol.* **82**, 10811–10819 (2008).
29. Luo, T., Douglas, J. L., Livingston, R. L. & Garcia, J. V. Infectivity enhancement by HIV-1 Nef is dependent on the pathway of virus entry: implications for HIV-based gene transfer systems. *Virology* **241**, 224–233 (1998).
30. Lai, R. P. *et al.* Nef decreases HIV-1 sensitivity to neutralizing antibodies that target the membrane-proximal external region of TMgp41. *PLoS Pathog.* **7**, e1002442 (2011).
31. Usami, Y. & Gottlinger, H. HIV-1 Nef responsiveness is determined by env variable regions involved in trimer association and correlates with neutralization sensitivity. *Cell Rep.* **5**, 802–812 (2013).
32. Campbell, E. M., Nunez, R. & Hope, T. J. Disruption of the actin cytoskeleton can complement the ability of Nef to enhance human immunodeficiency virus type 1 infectivity. *J. Virol.* **78**, 5745–5755 (2004).
33. Schaeffer, E., Gelezianas, R. & Greene, W. C. Human immunodeficiency virus type 1 Nef functions at the level of virus entry by enhancing cytoplasmic delivery of virions. *J. Virol.* **75**, 2993–3000 (2001).
34. Zhou, J. & Aiken, C. Nef enhances human immunodeficiency virus type 1 infectivity resulting from intervention fusion: evidence supporting a role for Nef at the virion envelope. *J. Virol.* **75**, 5851–5859 (2001).
35. Tobiume, M., Lineberger, J. E., Lundquist, C. A., Miller, M. D. & Aiken, C. Nef does not affect the efficiency of human immunodeficiency virus type 1 fusion with target cells. *J. Virol.* **77**, 10645–10650 (2003).
36. Cavois, M., Neideman, J., Yonemoto, W., Fenard, D. & Greene, W. C. HIV-1 virion fusion assay: uncoating not required and no effect of Nef on fusion. *Virology* **328**, 36–44 (2004).
37. Day, J. R., Munk, C. & Guatelli, J. C. The membrane-proximal tyrosine-based sorting signal of human immunodeficiency virus type 1 gp41 is required for optimal viral infectivity. *J. Virol.* **78**, 1069–1079 (2004).
38. Cavois, M., De Noronha, C. & Greene, W. C. A sensitive and specific enzyme-based assay detecting HIV-1 virion fusion in primary T lymphocytes. *Nature Biotechnol.* **20**, 1151–1154 (2002).
39. van der Aa, L. M. *et al.* A large new subset of TRIM genes highly diversified by duplication and positive selection in teleost fish. *BMC Biol.* **7**, 7 (2009).
40. Duggal, N. K. & Emerman, M. Evolutionary conflicts between viruses and restriction factors shape immunity. *Nature Rev. Immunol.* **12**, 687–695 (2012).
41. Briggs, J. A., Wilk, T., Welker, R., Kräusslich, H. G. & Fuller, S. D. Structural organization of authentic, mature HIV-1 virions and cores. *EMBO J.* **22**, 1707–1715 (2003).
42. Cohen, F. S. & Melikyan, G. B. The energetics of membrane fusion from binding, through hemifusion, pore formation, and pore enlargement. *J. Membr. Biol.* **199**, 1–14 (2004).
43. Razinkov, V. I. & Cohen, F. S. Sterols and sphingolipids strongly affect the growth of fusion pores induced by the hemagglutinin of influenza virus. *Biochemistry* **39**, 13462–13468 (2000).
44. Ciechonska, M. & Duncan, R. Lysophosphatidylcholine reversibly arrests pore expansion during syncytium formation mediated by diverse viral fusogens. *J. Virol.* **88**, 6528–6531 (2014).
45. Chen, A. *et al.* Fusion-pore expansion during syncytium formation is restricted by an actin network. *J. Cell Sci.* **121**, 3619–3628 (2008).

Acknowledgements We thank the Centre for AIDS Reagents, NIBSC, and NIH AIDS Research and Reference Reagent Program, Division of AIDS, for cell lines, plasmids and antibodies. We thank V. Adami and the CIBIO high-throughput screening and the Advanced Imaging facilities staff for technical assistance, G. De Silvestro, G. Mattiuzzo, C. Reinhard and L. Conti for reagents, G. Melikian, S. Basmaciogullari, P. Cherepanov, O. Fackler, N. Segata, F. Demicheli, A. Marcello, T. Fedrizzi and A. Helander for critical discussions. This work was supported by the Biotechnology Program of University of Trento, FP7 Marie Curie Career Integration grant number 322130 and Caritro 'Ricerca Biomedica' grant number 2013.0248 to M.P., National Institute of Health grant DP1DA034990 to J.L. and European Research Council grant 249968 to S.E.A.

Author Contributions A.R., A.C., S.Z., V.D.S., R.B., S.E.A., J.L., F.A.S. and M.P. designed the experiments. A.R., S.Z., A.C., V.D.S., R.B., S.L.G., S.M.M., A.N., F.A.S. and M.P. performed the experiments. All authors contributed to the assembly and writing of the manuscript. A.R., A.C. and S.Z. contributed equally to the study.

Author Information RNA-seq data have been deposited in NCBI Sequence Read Archive (SRA) under accession code SRP062444. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.P. (massimo.pizzato@unitn.it).

METHODS

Plasmids. *Env*-defective and *nef*-defective HIV-1^{NL4-3} have been described previously²². *Env*-defective and *glycoGag*-defective MLV were engineered to express GFP in place of *Env*. Unless otherwise indicated, single round HIV-1 *Env*-defective HIV-1 (NL4-3) was complemented with *Env* derived from HIV-1^{HXB2} expressed with the vector PBJ5 (ref. 22). Constructs for expression of other viral factors include: plasmids encoding *Env* from primary HIV-1 isolates (obtained from NIH AIDS Reagent Program); plasmids encoding wild-type and mutated HA-tagged *Nef* from HIV-1^{LA122} and *Nef* from primary HIV-1 isolates belonging to subtypes C, D, F; plasmids for expression of HA-tagged *Nef* from SIV^{mac} and SIV^{agm}²²; plasmids encoding untagged or HA-tagged MLV *glycoGag* truncated at residue 189 (ref. 16), pCAGGS expressing codon optimized Zaire Ebola virus glycoprotein (GenBank accession number KJ660346.2); pMDG⁴⁶ encoding VSV-G.

DNA encoding SERINC5 with or without the HA-tag were amplified from cDNA derived from JTAG cells. DNA sequence was confirmed to match the reference sequence with accession number NM_001174072.2. DNA encoding SERINC3 (reference sequence NM_006811) was custom synthesized (GeneWiz). For expression in mammalian cells, DNAs were cloned into expression vectors PCDNA3.1 (Life Technologies), PBJ5, PBJ6 (derived from PBJ5 by removing the SV40 origin of replication from the SV40-HTLV-1 hybrid promoter region), and pEGFPN1 (Clontech).

Increasing amount of SERINC5–HA expression in HEK293T cells was obtained by transfecting cells with PBJ6-, PBJ5-, and PCDNA3.1-based vector in increasing order (PBJ6 < PBJ5 < PCDNA3.1).

mRFP–Rab7 was a gift from A. Helenius (Addgene plasmid 14436). TagRFP657 was fused at the C terminus of *Nef* to generate pNef-Tag-RFP.

Cell lines. Cell lines used (also described in Extended Data Table 1 together with the source) were all tested for possible contamination with mycoplasma and tested negative. Cell line TE671 (Fig. 1a) is listed in the ICLAC database of commonly misidentified cell lines. However, for our purposes the nature of the cell line does not influence the outcome of the research which was only meant at investigating a correlation between the *Nef* requirement with gene expression. In addition to cell lines listed in Extended Data Table 1, TZM-bl indicator cells were obtained from the NIH AIDS Research and Reference Reagent program.

CRISPR–Cas9 knockout. Stable cell lines knocked out for SERINC5 were generated by transduction with LentiCRISPR (a gift from F. Zhang, Addgene plasmid 49535) after puromycin selection and, where indicated, clonal expansion. PX330 CRISPR–Cas9 (a gift from F. Zhang, Addgene plasmid 42230) was used for generating knockout by transient transfection, targeting simultaneously two different exons of the same gene. The following target sequences were used: 5′-GC TGAGGGACTGCCGAATCC-3′ (SERINC5-1, exon 2), 5′-GACGGTCCAC ATAGCGCC-3′ (SERINC5-2, exon 6), 5′-GGCGTACCACAGCTTGTTC-3′ (SERINC5-3, exon 8), 5′-GCATCGGCATAGCAACACG-3′ and 5′-CTATGC CGATGCTGTCTAG-3′ (SERINC1), 5′-CCGCATGTGCTTCGCCACGG-3′ and 5′-ATCCTGGTGGGCTCACCGT-3′ (SERINC2), 5′-ATAAATGAGGC GAGTACCG-3′ and 5′-CTCCGAGCGGAGTACACAA-3′ (SERINC3), 5′-TGATGACAGAAGCTTGTAGG-3′ and 5′-GGTTCCATTTTACTCAGGC C-3′ (SERINC4), 5′-GTGAACCGCATCGAGCTGAA-3′ (GFP).

To verify the occurrence of indels and the disruption of the SERINC5 open-reading frame (ORF) in clonal populations of JTAG cells stably transduced with the LentiCRISPR vector targeting SERINC5 exon 2 (using SERINC5-1 gRNA), genomic DNA was extracted from cells, a 228-nucleotide fragment encompassing exon 2 was amplified by PCR using primers 5′-TCGTCGGCAGCGTCAGATG TGTATAAGAGACAG-TAAGCAGATGCCTCTGTTCCTT-3′ and 5′-GTCT CGTGGGCTCGGAGATGTGTATAAGAGACAG-AATAGGACGAGCTGAAC ACGG-3′ (in which italic denotes the locus-specific sequence, and bold denotes the overhang adapters). A subsequent limited-cycle amplification step was performed to add multiplexing indices and Illumina sequencing adapters. Normalized and pooled libraries were, then, sequenced on the Illumina MiSeq system using v2 reagents (2 × 250-nucleotide paired-end reads).

Viruses and infectivity assay. Cell lines in Fig. 1a were infected with NL4-3 and NL4-3^{Nef} produced by transfection of HEK293T cells and transiently pseudotyped with VSV-G. Virus supernatant was collected 48 h after infection and inoculated onto TZM-bl cells in the presence of the protease inhibitor Saquinavir (10 μM) to limit infection to a single round of replication.

For all other experiments, virions limited to a single round of replication were used and were produced by transfection. JTAG and 174XCEM cells were transfected using electroporation, HT1080 using Mirus TransIT-2020, HEK293T cells by the calcium phosphate co-precipitation method. PBMC were transfected by nucleofection 48 h after stimulation with phytohaemagglutinin (PHA) and interleukin-2 (IL-2). As indicated, virus constructs were co-transfected together with other plasmids expressing *Env* glycoproteins, *Nef*, *glycoGag*, SERINC5, or PX330-based CRISPR–Cas9 vectors. Virus-containing culture supernatants were

collected 48 h after transfection, clarified by centrifugation at 300g for 5 min and passed through filters with 0.45-μm pores. Virus prepared in quadruplicate were then quantified using the SG-PERT reverse transcription assay⁴⁷, diluted three- or fivefold in a series of six steps and used to infect TZM-GFP reporter cells seeded one day before infection in 96-well plates. TZM-GFP is a modified version of TZM-bl containing an integrated nlsGFP reporter gene under the transcriptional control of the HIV-1 long terminal repeat. Infection of reporter cells was scored using the High Content Imaging System Operetta (Perkin Elmer) after counterstaining nuclei with Hoechst 33342 for each virus dilution. Those values falling into a linear dilution range (normally below 20% of infected cells) were used to calculate infectivity. Infectivity was calculated by dividing the number of infected cells in a well for the amount of reverse transcriptase activity associated to the virus inoculum, measured in mU⁴⁷.

Heterokaryons. Heterokaryons were produced following a strategy previously reported⁴⁸. Production of single round virions infectious only upon heterokaryon formation was obtained by transfecting one fusion partner with *env*-defective/*nef*-defective HIV-1^{NL4-3} and the other with PBJ5–HXB2–*Env*, PBJ5–*Nef*^{LA1} or the empty control vector PBJ5. To promote efficient fusion mediated by HIV-1 *Env*, plasmids encoding for CD4 and CXCR4 were co-transfected together with the *env*-defective provirus construct. Then 24 h after transfection, cells were co-cultured and progeny viruses collected 24 h later.

Preparation of RNA-seq libraries and sequencing. Five micrograms of total RNA extracted from seven highly *Nef*-dependent cell lines (JTAG, Jurkat E6.1, bl41, Ramos, CEM A301, CEM SS and HSB2) and eight low *Nef*-dependent cell lines (MT4, HT1080, RAJI, DAUDI, C8166, IMR90, CEMX174 and WI38) was subjected to rRNA depletion using Ribo-Zero Magnetic Gold Kit (Epicentre). RNA-seq libraries were prepared from the rRNA depleted RNAs extracted from the 15 cell lines (Fig. 1a) using a modified protocol of the Illumina TruSeq RNA Sample Prep Kit. Libraries were sequenced on the Illumina HiSeq 2000 using paired-end sequencing 2 × 100 bp. Raw reads were mapped against the human (hg19) genome reference using tophat2 (ref. 49). RPM³⁰ values were estimated for each transcript in each sample with a custom pipeline. Genes were ranked according to Pearson correlation between their relative expression (RPM) in cell line and the corresponding *Nef*⁺/*Nef*[−] infectivity ratio (Fig. 1a). The computations were performed at the Vital-IT Center (<http://www.vital-it.ch>) for high-performance computing of the SIB Swiss Institute of Bioinformatics in Geneva.

Microscopy. JTAG cells were electroporated with constructs expressing *Nef*-TagRFP657 or the control TagRFP657, *Nef*–HA, HA–glycoMa, SERINC5–GFP and Rab7–RFP as indicated. Then 48 h after transfection, cells were overlaid on poly-L-lysine coated glass slides, fixed with 4% paraformaldehyde and permeabilized with 0.1% Triton X-100. The HA tag was detected by staining with mouse anti-HA (HA.11, Covance) and the secondary antibody Alexa 633 (Life Technologies). Images were acquired using a Leica TCS SP5 confocal microscope.

Western blotting. Cell lysates and virion pellets were analysed by SDS–PAGE and western blotting. In brief, viral particles were collected 48 h after transfection, centrifuged at 300g to remove cell debris and filtered. The clarified supernatants were overlaid on 25% sucrose cushion and concentrated at 100,000g. The pellets were resuspended directly in Laemmli buffer (supplemented with 50 mM TCEP), normalized by reverse transcriptase assay and resolved by SDS–PAGE. After observing that SERINC5 and SERINC3 form aggregates that are lost while clarifying the cell lysate or fail to enter the separating gel, cells were lysed directly in Laemmli buffer containing TCEP (Sigma, final concentration 50 mM, pH 7.0) and avoiding boiling. Samples were loaded on gel after a 5-pulse sonication. Having failed to find a commercially available antibody capable of detecting the endogenous protein, probing was performed using mouse anti-HA (HA.11, Clone 16B12, Covance), mouse or rabbit anti-β-actin (Li-COR), anti-HIV-1 p55/p24 (National Biological Standards Board), anti-gp41 Chessie-8 (obtained from the NIH AIDS Reagent Program, Division of AIDS, NIAID, NIH from G. Lewis), mouse anti-Cre (Mab3120, Chemicon) and secondary antibodies IRDye 680 and IRDye 800 (Li-COR). Blots were imaged using an Odyssey Imager system (Li-COR).

nlsCre delivery assay. A packaging vector based on p8.9 lentiviral gag-pol expressing plasmid⁴⁶ (8.9-Cre) was generated to carry an insertion of nlsCre between MA and CA flanked by native HIV-protease cleavage sites for processing and release of proteins from Gag.

A Cre-responsive nuclear RFP-expressing lentiviral vector (p-lenti LoxP-Blasti-mRFP) was created. It consists of a nls-mRFP sequence lacking the translation initiation codon and preceded by a sequence encoding the blasticidin antibiotic resistance (Bla) between two *loxP* sites. The nlsRFP is translationally inactive unless Cre-mediated recombination of *loxP* and excision of Bla occurs, providing an authentic translation initiation for mRFP. A TZM-bl-GFP derivative cell line (TZM-GFP) stably transduced with p-lenti LoxP-Blasti mRFP was generated (TZM-GFP-LoxP-RFP) to detect delivery of nlsCre, and Tat-driven expression of nlsGFP.

To package nlsCre in retrovirus particles, HIV-1 was produced by mixing 8.9-Cre together with the *env*-defective (and *nef*-defective where applicable) NL4-3 provirus at a ratio of 1:2. Virus was produced by cotransfecting HEK293T cells with the viral constructs together with PBj5-HXB2-*env* or vectors for expression of VSV-G and Ebola glycoprotein, and plasmids encoding SERINC5 or the empty vector. To achieve increasing level of expression, SERINC5 was expressed from vector PBj6, PBj5 and PCDNA3.1 (in increasing order). Progeny virus was inoculated onto TZM-GFP-LoxP-RFP and red and green fluorescence quantified 48 h later using the High Content Imaging System Operetta (Perkin Elmer) after counterstaining nuclei with Hoechst 33342, following the method described for infectivity.

BLAM-VpR assay. Virus was produced by transfection of HEK293T with the calcium phosphate method in 10 cm tissue culture plates with 10 µg of NL4-3 *Envfs/Nefts* (bearing a frameshift) together with 2 µg HIV-1 *Env* expressor, 5 µg of BLAM-VpR vector³⁸ and 5 µg of SERINC5 expression vectors or the empty vector control.

Target cells (TZM-bl) were seeded in clear bottom 96-well plates (Optiplates, Perkin Elmer) at a density of 25,000 cells per well in phenol-Red-free medium one day before assay. Virus samples were normalized for reverse transcriptase activity content and added to wells (200 µl) serially diluted as described for infectivity. Cells were spin-infected for 2 h at 4 °C at 1,550g, virus was removed, cells washed twice with complete medium and incubated for 90 min at 37 °C. Medium was then replaced with GeneBlazer substrate loading solution containing 2 µM CCF2AM (GeneBlazer *In vivo* Detection Kit, Life Technologies) and 2.5 mM Probenecid (Sigma). Cells were incubated overnight at 11 °C, fixed with 2% paraformaldehyde and plates analysed using the Operetta imaging system for blue and green fluorescence to reveal the number of blue positive cells. Transduction units were derived from the number of blue positive cells divided per reverse transcriptase activity associated to the virus inocula as described for infectivity.

Quantification of HIV-1 reverse transcription products. NL4-3 normalized based on reverse transcriptase activity was incubated with target cells (NP2-CD4-CXCR4). Cell-free virions were normalized by reverse transcriptase activity and incubated with target cells in 6-well plates for 12 h. For each virus, infections were also performed in the presence of 40 µM AZT, to control for contamination of plasmid DNA in the PCR reaction. Cells were collected and washed extensively with PBS. Total DNA was extracted (Qiagen, Qiaamp DNA mini kit), quantified, and subjected to real-time PCR with a Biorad CFX96 cyclor. cDNA was detected with SYBR-Green I based reactions using 100 ng template DNA and 320 nM of each primer pair (5'-ACAAGCTAGTACCAGTTGAGCCAGATAAG-3' and 5'-GCCGTGCGCGCTTCAGCAAGC-3') in 20 mM Tris-Cl, pH 8.3, 5 mM (NH₄)₂SO₄, 20 mM KCl, 5 mM MgCl₂, 0.1 mg ml⁻¹ BSA, 1/20,000 SYBR Green I (Sigma), and 200 µM dNTPs. The PCR was programmed for 40 cycles of denaturation at 95 °C for 5 s, annealing 55 °C for 5 s, extension at 72 °C for 20 s and acquisition at 80 °C for 5 s. Relative quantification of retroviral cDNA sequences was obtained with respect to standard curves prepared from serial dilutions of DNA derived from the cell culture with the highest infection, diluted in DNA extracted from non-infected cells.

PBMC. Buffy coats obtained from anonymous blood donors were provided by the Department of Immunotransfusion, Padova University Hospital, for experiments involving virus production, or purchased from the New York Blood Center, for interferon induction studies. PBMC were isolated using Ficoll-Paque Plus (GE Healthcare).

Isolation, stimulation and treatment of dendritic cells. CD14⁺ monocytes were enriched from PBMC by positive selection using CD14 MicroBeads following the manufacturer's protocol (Miltenyi Biotec). CD14⁺-enriched cell populations were counted, centrifuged at 200g for 10 min, and resuspended at 2 × 10⁶ cells ml⁻¹ in RPMI-1640 supplemented with 5% human AB+ serum, 1 × MEM non-essential

amino acids (NEAA), 20 mM L-glutamine, 25 mM HEPES, 1 mM sodium pyruvate and 50 µM β-mercaptoethanol. To induce differentiation of monocytes into dendritic cells, cells were cultured for 5 days in GM-CSF (50 ng ml⁻¹) and IL-4 (25 ng ml⁻¹), both cytokines from R&D Systems. Dendritic cells were treated with LPS (100 ng ml⁻¹, LPS-EK Ultrapure, Invivogen) or IFN-β (37 ng ml⁻¹, PBL Assay Science). Cells were collected at various time points (*t* = 0 h, 2 h, 6 h, 24 h) after the LPS and IFN-β treatments for RNA extraction and subsequent RT-PCR analysis.

Isolation, stimulation and treatment of CD4⁺ T cells. CD4⁺ T cells were isolated from CD14-depleted PBMCs by positive selection using CD4 magnetic microbeads (Miltenyi Biotec) and plated at 2 × 10⁶ cells ml⁻¹ in RPMI-1640, supplemented with 10% FBS, 25 mM HEPES, 1 mM sodium pyruvate, 1 × MEM NEAA, and 1 × GlutaMAX (Life Technologies). In one experiment, CD4⁺ T cells were treated directly with LPS (100 ng ml⁻¹) or IFN-β (37 ng ml⁻¹, PBL Assay Science). Separately, CD4⁺ T cells from the same donors were stimulated with 4 µg ml⁻¹ of PHA-M for 48 h, 20 IU ml⁻¹ IL-2 was added, and cells were stimulated with LPS or IFN-β. Cells were collected at various time points (*t* = 0 h, 2 h, 6 h, 24 h) after the LPS and IFN-β treatments for RNA extraction and subsequent RT-PCR analysis. Jurkat T cells were cultured and stimulated similarly.

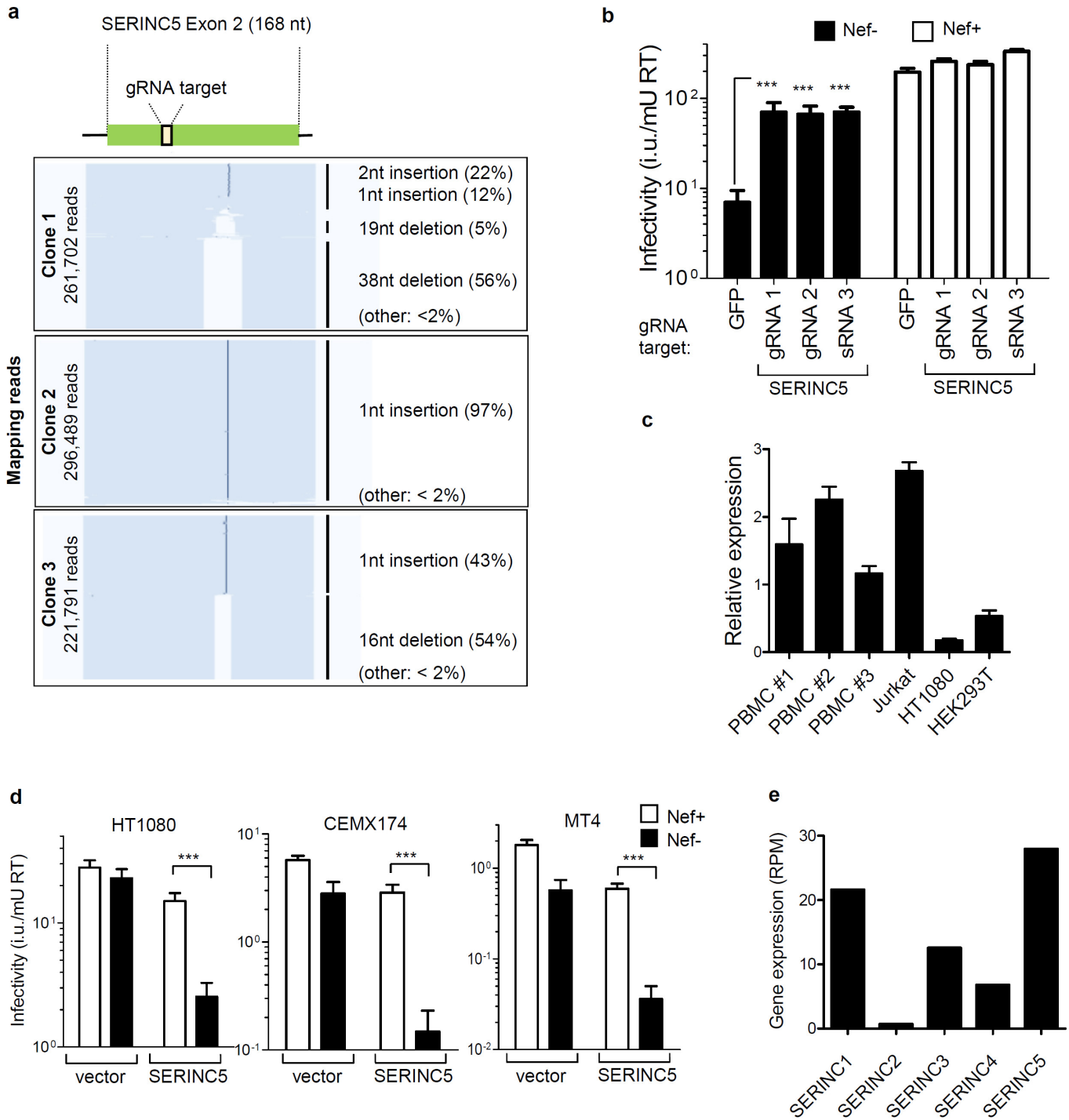
RNA isolation and qRT-PCR. RNA was isolated using RNeasy Plus Mini Kit (Qiagen 74134) with additional on column DNase treatment (Qiagen 79254) and reverse transcribed with SuperScript VILO Master Mix (Invitrogen 11755050). Gene expression was assayed on a Biorad CFX96 Real-Time PCR detection system.

For quantification of SERINC5 transcripts in cell lines and PBMC (Extended Data Fig. 2c), the SYBR-Green-based real-time PCR method was used with the following primers 5'-TAAGCAGATGCCTTCTGTTCCCTT-3' and 5'-AATAG GACGAGCTGAACACGG-3' (for *SERINC5*) and 5'-GACAGGATGCAGAAG GAGATTACTG-3' and 5'-CTCAGGAGGAGCAATGATCTTGAT-3' (for β-actin used as normalization control).

For Extended Data Fig. 4, gene expression was measured using TaqMan Gene Expression Master Mix (Life Technologies 4369016) and the following TaqMan probes and primers sets: *SERINC5* (Hs00968169_m1, Life Technologies 4351372) and *SERINC3* (Hs01566572_m1), *CXCL10* (Hs00171042_m1) and, as a normalization control, *OAZ1* (Hs00427923_m1, Life Technologies 4331182).

Statistics. Statistical tests were performed using GraphPad Prism. Given the nature of the experiments and the type of samples, significance of differences was assessed with unpaired two-tailed Student's *t*-test. Variance was estimated by calculating the standard deviation in each group, as represented by error bars. Variances between groups of samples were compared using the F-test function integrated in GraphPad. No statistical methods were used to predetermine sample size. Unless otherwise specified in figure legends, all experiments were performed independently at least three times and 'n' indicates technical replicates, with a representative experiment being shown. Experiments were not randomized, and the investigators were not blinded to allocation during experiments and outcome assessment.

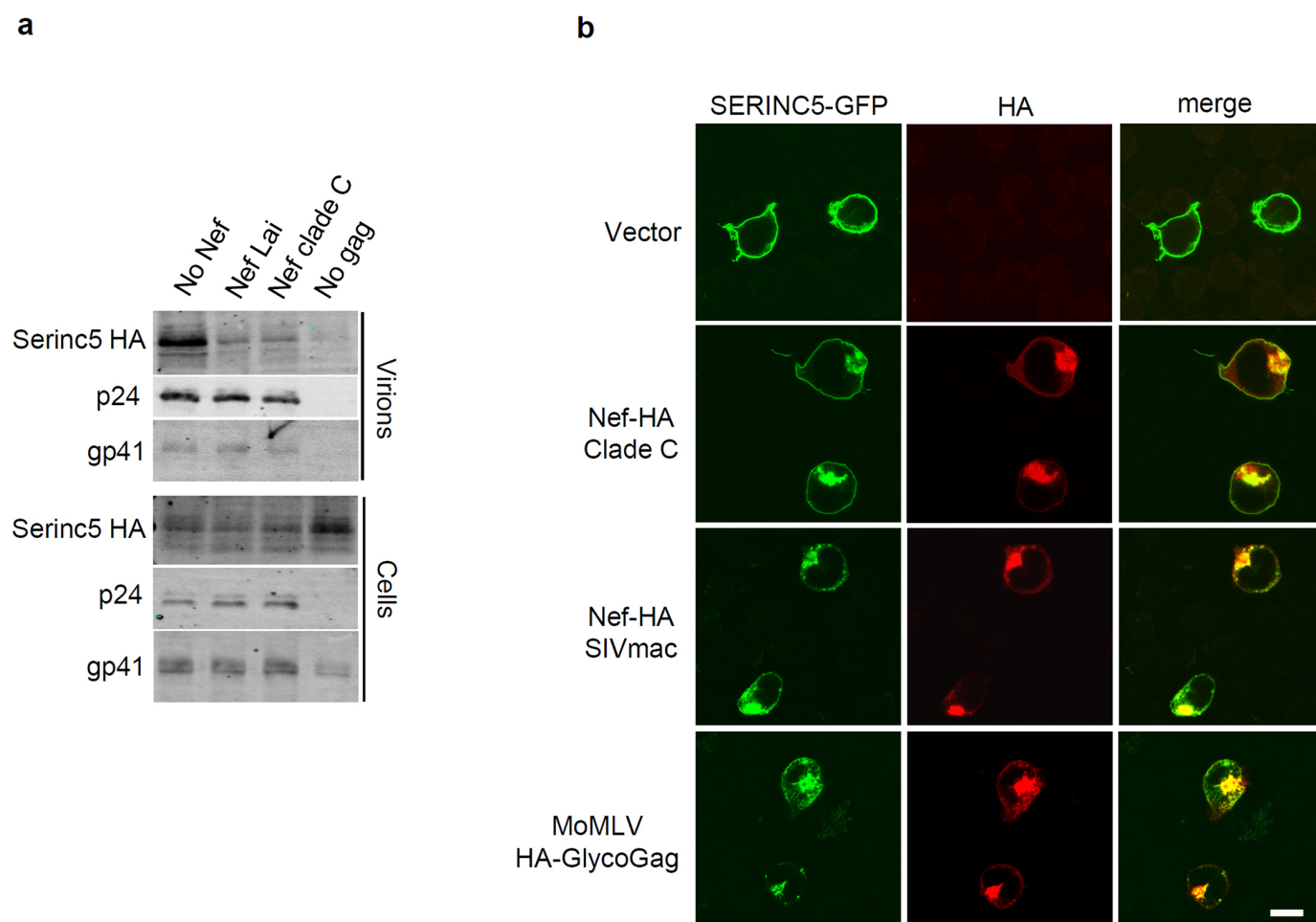
46. Zufferey, R., Nagy, D., Mandel, R. J., Naldini, L. & Trono, D. Multiply attenuated lentiviral vector achieves efficient gene delivery *in vivo*. *Nature Biotechnol.* **15**, 871–875 (1997).
47. Pizzato, M. *et al.* A one-step SYBR Green I-based product-enhanced reverse transcriptase assay for the quantitation of retroviruses in cell culture supernatants. *J. Virol. Methods* **156**, 1–7 (2009).
48. Simon, J. H., Gaddis, N. C., Fouchier, R. A. & Malim, M. H. Evidence for a newly discovered cellular anti-HIV-1 phenotype. *Nature Med.* **4**, 1397–1400 (1998).
49. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
50. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5**, 621–628 (2008).



Extended Data Figure 1 | SERINC5 is an inhibitor of HIV-1 infectivity.

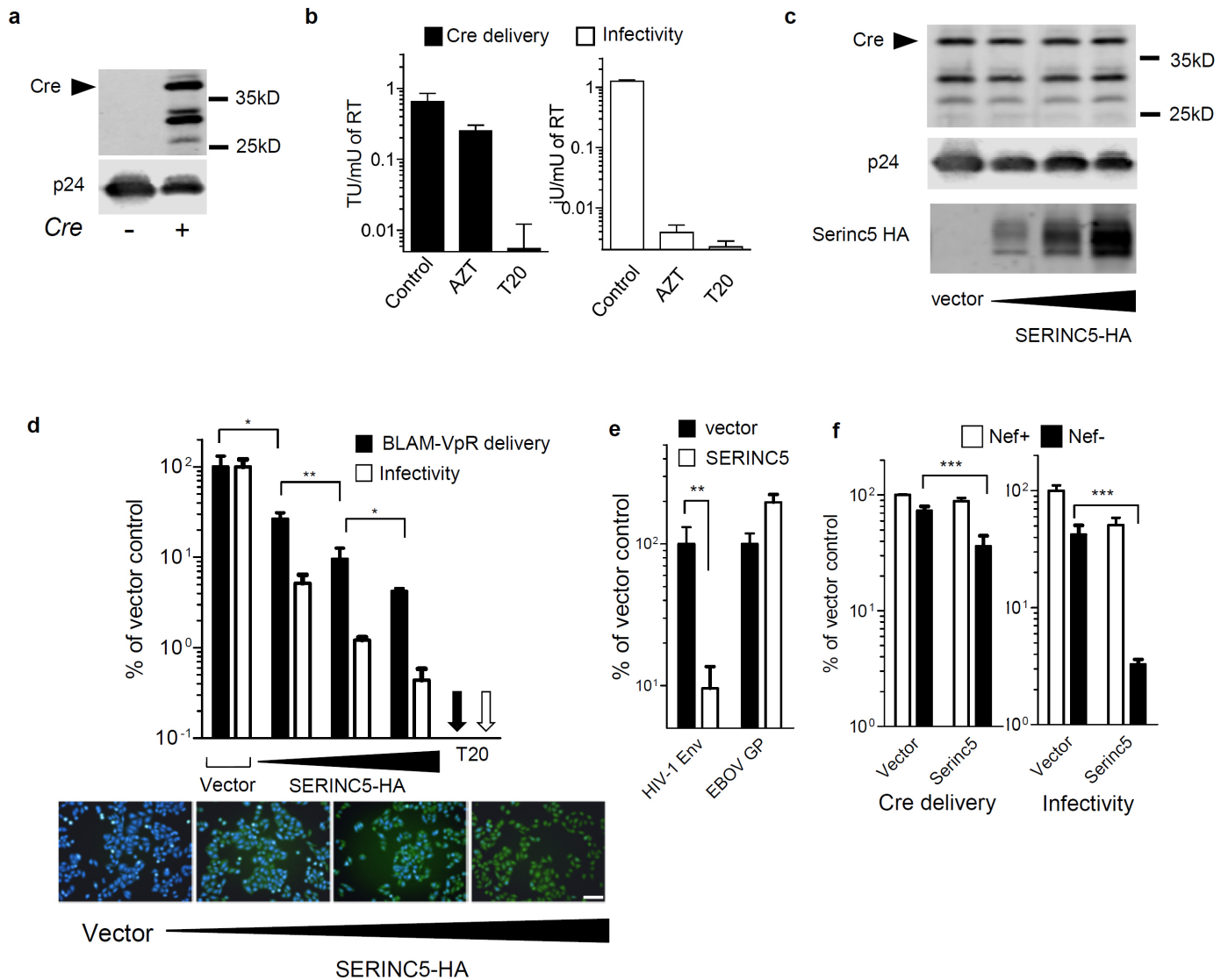
a, Mapping of the INDELS in the genomic locus spanning SERINC5 exon 2 in JTAG cell clonal populations from Fig. 2a. **b**, Infectivity of HIV-1 from JTAG cells stably transduced with lentiCRISPR targeting GFP or SERINC5 in three different exons ($n = 4$, experiment replicated twice). **c**, Relative expression of SERINC5 in primary cells and in cell lines measured by qPCR

normalized by expression of *ACTB* ($n = 3$). **d**, Infectivity of HIV-1 from the indicated cell lines expressing SERINC5 ($n = 4$, experiments were replicated twice). Mean \pm s.d., unpaired two-tailed *t*-test, *** $P < 0.001$. **e**, Expression levels of the five SERINC genes in JTAG cells obtained from RNA-seq.



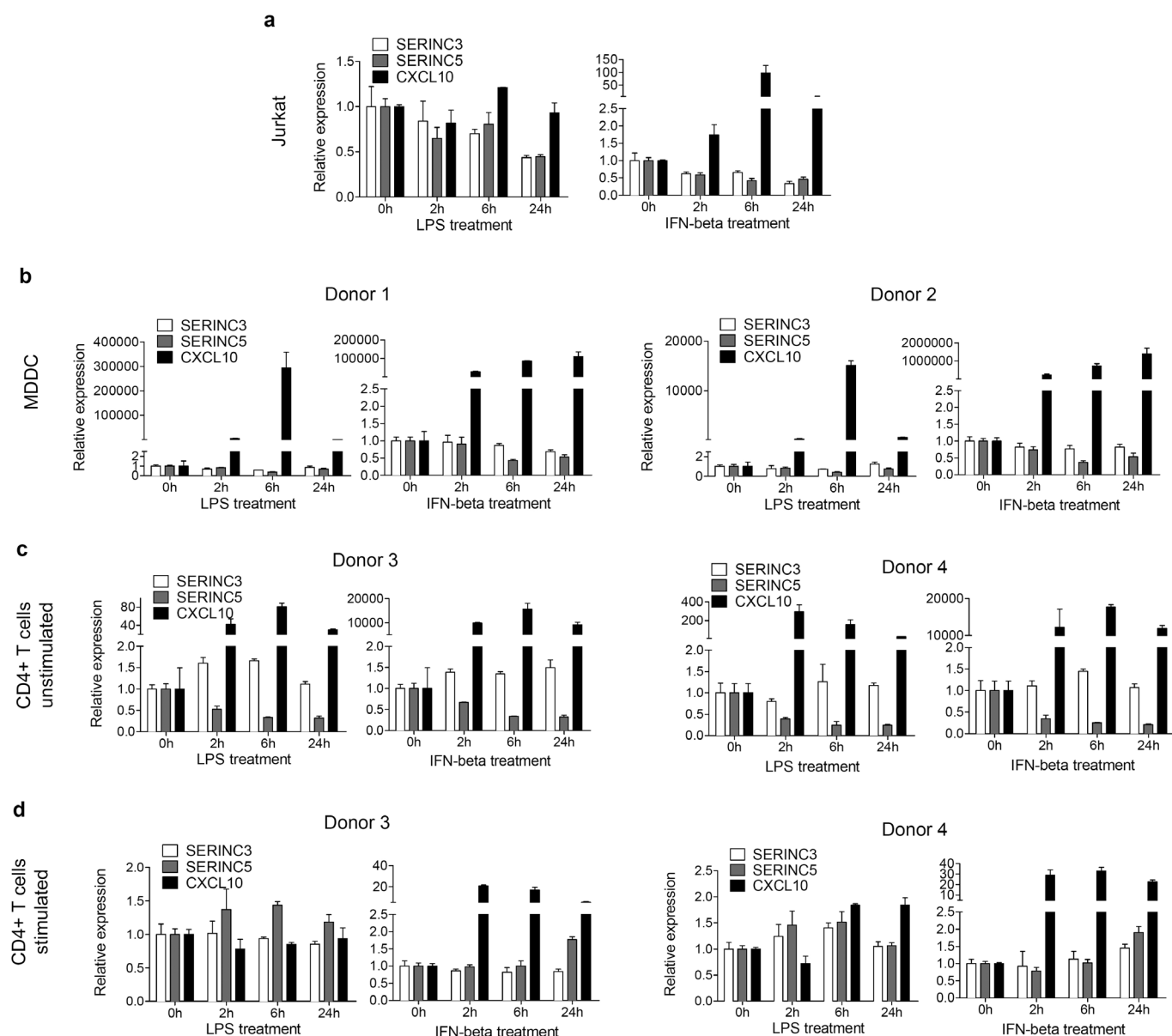
Extended Data Figure 2 | Nef and glycoGag expression result in relocalization of SERINC5 to an endosomal compartment and prevent its incorporation into virions. **a**, Single round Nef-defective NL4-3 produced by cotransfection of HEK293T cells with plasmids expressing Nef proteins or the empty vector control, and PBj6-SERINC5-HA: immunoblotting of

virions and cell lysates from producer cells. **b**, Immunofluorescence staining of JTAG cells transfected to express SERINC5-GFP, Nef-HA from HIV-1 isolate 97ZA012 (clade C), from SIV^{mac239}, HA-glycoGag or an empty vector control. Scale bar, 10 μm .



Extended Data Figure 3 | SERINC5 inhibits cytoplasmic delivery of virion content. **a**, Immunodetection of Cre-recombinase (38 kDa) and p24 in HIV-1 particles. **b**, Effect of 1 μ M AZT or 100 nM T20 on Cre-delivery and virus infectivity (TU, transducing units). **c**, Immunoblotting of HIV-1 virus particles produced from HEK293T expressing increasing levels of SERINC5-HA. **d**, Effect of SERINC5 on virus fusion measured with BLAM assay T20 served as

a negative control. ($n = 4$, experiment replicated twice). **e**, Cre delivery by EBOV-GP pseudotyped HIV-1 particles. **f**, Inhibition of Cre delivery and counteraction by Nef on HIV-1 from HEK293T expressing SERINC5. Mean \pm s.d., $n = 4$, unpaired two-tailed t -test, $*P < 0.05$, $**P < 0.01$, $***P < 0.001$. Scale bar, 100 μ m.



Extended Data Figure 4 | SERINC3 and SERINC5 expression is not induced by interferon nor LPS treatments. a–d, Relative gene expression levels of *SERINC3*, *SERINC5* and *CXCL10* in response to treatment with IFN- β and LPS in Jurkat (a), monocyte-derived dendritic cells from two donors (MDDC, b),

CD4⁺ primary T cells unstimulated (c) or stimulated with PHA (d) from two donors. Expression of the housekeeping gene *OAZ1* was used as a normalization control. Mean \pm s.d., $n = 3$.

Extended Data Table 1 | Description of the cells lines used in Fig. 1a

| Cell Line | Cell Type | Source |
|-------------|---|---|
| Jurkat E6.1 | T Lymphocyte, Acute T Cell Leukemia | ATCC |
| Jurkat TAg | T Lymphocyte, Acute T Cell Leukemia. Derivative of Jurkat E6.1, contains Sv40 LargeT antigen | Heinrich Gottlinger, DFCI, Harvard University |
| bl41 | B-Lymphocyte, Burkitt's lymphoma | Paul Farell, Imperial College London |
| Ramos | B-Lymphocyte, Burkitt's lymphoma | ATCC |
| CEM-CCRF | T-Lymphocyte, Acute Lymphoblastic Leukemia | ATCC |
| CEM/A3.01 | T-Lymphocyte, Acute Lymphoblastic Leukemia Derivative of CEM-CCRF | NIH AIDS Reagent Program |
| CEMSS | T-Lymphocyte, Acute Lymphoblastic Leukemia. | NIH AIDS Reagent Program |
| HSB-2 | T- Lymphocyte, Acute Lymphoblastic Leukemia, CD4- | NIH AIDS Reagent Program |
| H9 | T-Lymphocyte, human cutaneous T cell lymphoma. Derivative of HUT-78 | NIH AIDS Reagent Program |
| DG7 HAD | B-Lymphocyte, Burkitt's lymphoma | Sidney Grossberg, University of Wisconsin |
| Hela | Epithelial, cervix adenocarcinoma | ATCC |
| Akata | B-Lymphocyte, Burkitt's lymphoma | Paul Farell, Imperial College London |
| SupT1 | T-Lymphocyte, T-Cell Lymphoblastic Lymphoma | NIH AIDS Reagent Program |
| A549 | Epithelial, lung carcinoma | ATCC |
| HepG2 | Hepatocellular Carcinoma | ATCC |
| HCT116 | Epithelial, colorectal carcinoma | ATCC |
| MCF7 | Epithelial, adenocarcinoma | ATCC |
| HUT78 | T-Lymphocyte, human cutaneous T cell lymphoma | NIH AIDS Reagent Program |
| 293T | Epithelial, embryonic Kidney | ECACC |
| MT2 | T Lymphocyte, T-cell leukemia | NIH AIDS Reagent Program |
| TE671 | Rhabdomyosarcoma | Yasuhiro Takeuchi, UCL, London |
| Huh-7 | Hepatocellular Carcinoma | Michel Strubin, University of Geneva |
| LL24 | Lung Fibroblast | ATCC |
| RAJI | B-Lymphocyte, Burkitt's lymphoma | ATCC |
| C8166 | T lymphocytes, T cell leukaemia | NIH AIDS Reagent Program |
| WI38 | Lung fibroblast | ATCC |
| Daudi | B-Lymphocyte, Burkitt's lymphoma | Paul Farell, Imperial College London |
| MT4 | T lymphocytes, T cell leukaemia | NIH AIDS Reagent Program |
| IMR-90 | Lung fibroblast | ATCC |
| CEMX174 | Lymphocytes, Fusion between a B cell line and a human T cell line | NIH AIDS Reagent Program |
| HT1080 | Epithelial, fibrosarcoma | Yasuhiro Takeuchi, UCL, London |

SERINC3 and SERINC5 restrict HIV-1 infectivity and are counteracted by Nef

Yoshiko Usami^{1*}, Yuanfei Wu^{1*} & Heinrich G. Göttlinger¹

HIV-1 Nef and the unrelated mouse leukaemia virus glycosylated Gag (glycoGag) strongly enhance the infectivity of HIV-1 virions produced in certain cell types in a clathrin-dependent manner. Here we show that Nef and glycoGag prevent the incorporation of the multipass transmembrane proteins serine incorporator 3 (SERINC3) and SERINC5 into HIV-1 virions to an extent that correlates with infectivity enhancement. Silencing of both SERINC3 and SERINC5 precisely phenocopied the effects of Nef and glycoGag on HIV-1 infectivity. The infectivity of *nef*-deficient virions increased more than 100-fold when produced in double-knockout human CD4⁺ T cells that lack both SERINC3 and SERINC5, and re-expression experiments confirmed that the absence of SERINC3 and SERINC5 accounted for the infectivity enhancement. Furthermore, SERINC3 and SERINC5 together restricted HIV-1 replication, and this restriction was evaded by Nef. SERINC3 and SERINC5 are highly expressed in primary human HIV-1 target cells, and inhibiting their downregulation by Nef is a potential strategy to combat HIV/AIDS.

Nef is an accessory protein encoded by HIV-1 and other primate lentiviruses. *In vivo*, Nef is a major pathogenicity determinant that is required for high virus loads^{1–3}. Although not essential for virus replication in cell culture, Nef enhances virus spreading in primary CD4⁺ T cells, particularly when such cells are infected before mitogenic stimulation^{4–6}. Nef robustly downregulates the viral entry receptor CD4 from the surface of virus-producing cells by inducing its clathrin-dependent endocytosis and subsequent lysosomal degradation^{7–10}. A recent study suggests that an important physiological function of CD4 downregulation by Nef is to prevent the CD4-induced exposure of epitopes in HIV-1 envelope (Env) that make infected cells susceptible to antibody-dependent cell-mediated cytotoxicity¹¹. Apart from CD4, Nef downregulates several other cell surface proteins¹². The selective down-modulation of HLA-A and HLA-B but not of HLA-C by Nef serves to protect infected cells both from cytotoxic T cells and from natural killer cells^{13–15}. The Nef proteins of most primate lentiviruses also down-modulate the T cell receptor complex, which is thought to protect infected T cells from activation-induced cell death in non-pathogenic natural SIV infections¹⁶. This function of Nef was lost in HIV-1 and closely related viruses, which may contribute to the pathogenicity of HIV-1 in humans¹⁶.

One of the most conserved yet poorly understood functions of Nef is the enhancement of progeny virion infectivity^{17,18}. Although Nef exerts its effect on HIV-1 infectivity in virus producer cells, it does not detectably affect virus morphogenesis or maturation^{19–22}. Nevertheless, progeny virions produced in the absence of Nef do not efficiently reverse transcribe their genome in target cells^{19,20}. It has been reported that high levels of cell-surface CD4 inhibit the release or infectivity of HIV-1 progeny virions, and that Nef relieves these effects^{23,24}. However, the enhancement of HIV-1 infectivity depends on residues within Nef that are dispensable for its ability to downregulate CD4 (ref. 25). Furthermore, Nef enhances HIV-1 progeny virion infectivity even when CD4 is not expressed or cannot be downregulated^{17,19,20}. Finally, the glycoGag protein of Moloney murine leukaemia virus (MLV) closely mimics the effect of Nef on HIV-1 infectivity, even though glycoGag does not downregulate CD4

(ref. 26). MLV glycoGag is an accessory protein whose translation begins at an inefficient CTG start codon upstream and in-frame with the *gag* gene²⁷. The resulting product is a type II transmembrane protein with an amino-terminal cytosolic non-Gag portion and an extracellular Gag domain²⁸. The potent Nef-like activity of glycoGag on HIV-1 infectivity resides entirely in its cytosolic domain, which is unrelated to Nef²⁹. Nevertheless, the effects of Nef and glycoGag on HIV-1 infectivity appear mechanistically related. Both are similarly dependent on the producer cell type²⁶, are similarly determined by variable regions of HIV-1 Env³⁰, and exhibit a similar reliance on clathrin-mediated endocytosis^{29,31,32}. However, the molecular basis for these similarities remains unknown.

Nef inhibits the incorporation of SERINC proteins

Because of the essential role of the endocytic machinery in the enhancement of HIV-1 infectivity by Nef or glycoGag, we examined the possibility that both proteins downregulate a restriction factor that gets incorporated into assembling virions in their absence. To identify factors whose incorporation is prevented by both Nef and glycoGag, we conducted a proteomic analysis of OptiPrep gradient-purified virions produced by T lymphoid cells infected with wild-type (Nef⁺) or Nef⁻ HIV-1_{NL43}, or with a version that encodes a fully active minimal glycoGag (termed glycoMA³⁰) instead of Nef (Extended Data Fig. 1a). The only host protein that could reproducibly be identified in Nef⁻ virion samples in independent experiments but was not identified in any Nef⁺ or glycoMA virion sample was SERINC3, a member of a family of putative carrier proteins with at least 10 transmembrane domains³³ (Extended Data Fig. 1b). In one experiment, STOM and PFKP were also identified in Nef⁻ but not in Nef⁺ or glycoMA virion samples (Extended Data Fig. 1b). However, in another experiment, STOM was identified in all virions samples, and PFKP was not identified in any sample. Thus, STOM and PFKP were not further pursued. Immunoblotting of virion samples confirmed that the incorporation of haemagglutinin (HA)-tagged SERINC3 is strongly inhibited by the Nef proteins of several laboratory-adapted and primary HIV-1 isolates from different clades

¹Department of Molecular, Cell and Cancer Biology, University of Massachusetts Medical School, Worcester, Massachusetts 01605, USA.

*These authors contributed equally to this work.

(Fig. 1a) and by glycoMA (Extended Data Fig. 2a). Furthermore, the effects of glycoMA truncation mutants on the incorporation of SERINC3–HA (Extended Data Fig. 2a) correlated closely with their abilities to enhance HIV-1 infectivity²⁹. Two of the Nef proteins tested did not inhibit the incorporation of SERINC3–HA (Fig. 1a), and one of these (Nef_{90CF056}) also had no effect on HIV-1 infectivity (Fig. 1c). Because the other (Nef_{SF2}) did enhance HIV-1 infectivity (Fig. 1c), we examined its effect on the incorporation of other human SERINC family members. Although Nef_{SF2} did not affect the incorporation of SERINC3–HA (Fig. 1a), it strongly inhibited the incorporation of SERINC5–HA (Fig. 1b). Among the primary Nef proteins examined, those that were most active in enhancing HIV-1 infectivity (Nef_{97ZA012} and Nef_{93BR020}) strongly inhibited the incorporation of both SERINC3 and SERINC5, the less active Nef_{94UG114} was a less effective inhibitor particularly of SERINC5 incorporation, and the inactive Nef_{90CF056} inhibited neither SERINC3 nor SERINC5 incorporation (Fig. 1a–c). Like the most active Nef proteins, wild-type glycoMA, which enhances HIV-1 infectivity at least as potently³⁰, also strongly inhibited the incorporation of both SERINC3 and SERINC5 (Extended Data Fig. 2a, b). Furthermore, the effects of glycoMA truncation mutants on SERINC5 incorporation (Extended Data Fig. 2b), like those on SERINC3 incorporation (Extended Data Fig. 2a), correlated with their effects on HIV-1 infectivity enhancement²⁹.

Subcellular localization of SERINC5

SERINC5–mCherry clearly localized to the plasma membrane and to filopodia-like protrusions when expressed alone, but accumulated in perinuclear vesicles when co-expressed with Nef or glycoGag (Extended Data Fig. 3a and data not shown). Furthermore, SERINC5(iHA), which contains an internal HA tag next to a conserved consensus glycosylation site within a proposed extracellular loop³⁴, could be readily detected on the surface of transfected Jurkat TAg (JTAG) T lymphoid cells by flow cytometry, and its surface expression was greatly reduced when either Nef_{SF2} or glycoGag were co-expressed (Extended Data Fig. 3b). We infer that Nef and glycoGag decrease the virion-association of SERINC5 by decreasing its cell surface levels.

Effects of exogenous SERINC5 on HIV infectivity

HIV-1 virions produced in Jurkat T lymphoid cells are more dependent on Nef for optimal infectivity than virions produced in 293T cells²⁶, which in turn are more dependent on Nef than virions produced in exceptionally permissive MT4 cells³⁵. Interestingly, the relative requirement for Nef correlates with SERINC5 messenger RNA levels, which are high in Jurkat cells, lower in 293T cells, and lower yet in MT4 cells (Extended Data Fig. 4a, b). SERINC5 mRNA levels in unstimulated or stimulated human peripheral blood mononuclear

cells (PBMC) were even higher than in Jurkat cells, and were not further increased by treatment with interferon- α (INF- α) (Extended Data Fig. 4b, c).

In single cycle replication assays, exogenous SERINC5 reduced the specific infectivity of Nef[−] HIV-1 virions produced in 293T cells for TZM-bl indicator target cells >100-fold, even when as little as 100 ng of the relatively weak pBJ5-based SERINC5 expression vector was used (Fig. 2a). Under the same conditions, exogenous SERINC3 reduced progeny virus infectivity only two- to threefold (Fig. 2a). However, although endogenous SERINC5 mRNA levels in 293T cells are low, these cells have relatively high endogenous SERINC3 mRNA levels (Extended Data Fig. 4b). Even at 500 ng, the vectors expressing SERINC3 or SERINC5 did not affect virus particle production, Gag processing, or HIV-1 Env incorporation (Fig. 2b). However, late reverse transcriptase products in target cells exposed to Nef[−] HIV-1 virions produced in 293T cells transfected with 500 ng of the SERINC5 expression vector were reduced >100-fold (Fig. 2c).

To examine whether SERINC5 affects HIV-1 virion fusion with target cells, we co-expressed a chimaeric β -lactamase-Vpr (BlaM-Vpr) protein that is taken up into virions³⁶. Fusion was then quantified based on the cleavage of a fluorescent substrate after the transfer of BlaM-Vpr from virions into target cells. We initially used 1 μ g of the SERINC5 expression vector to compensate for potential competition by the strong promoter driving BlaM-Vpr expression, and found that the ability of Nef[−] HIV-1 progeny virions to fuse with target cells was largely abolished (Extended Data Fig. 5). However, 100 ng of the SERINC5 expression vector, which reduced the infectivity of Nef[−] HIV-1 virions ~20-fold when co-transfected with the BlaM-Vpr expression vector, caused only an ~4-fold reduction in the ability to fuse with target cells (Extended Data Fig. 5).

The effects on HIV-1 infectivity were specific, because 500 ng of the vectors expressing SERINC3 or SERINC5 had at most modest effects on the infectivities of Env[−] HIV-1 particles pseudotyped with the vesicular stomatitis virus G protein (VSV-G) (Fig. 2d), which do not require Nef or glycoGag for optimal infectivity^{26,37,38}. Surprisingly, the incorporation of HA-tagged SERINC5 into Env[−] HIV-1 particles was reduced in the presence of VSV-G (Fig. 2e). Thus, reduced incorporation may have contributed to the relative resistance of VSV-G-pseudotyped HIV-1 to exogenous SERINC5. Crucially, the effect of exogenous SERINC5 on Nef[−] HIV-1 was counteracted by Nef_{SF2} or glycoGag expressed *in trans* (Fig. 2f). Indeed, exogenous SERINC5 had no effect whatsoever in the presence of glycoGag (Fig. 2f). In two independent experiments performed with cells from different donors, exogenous SERINC5 also greatly reduced the infectivity of Nef[−] HIV-1 virions produced in 293T cells for primary human target cells (Extended Data Fig. 6).

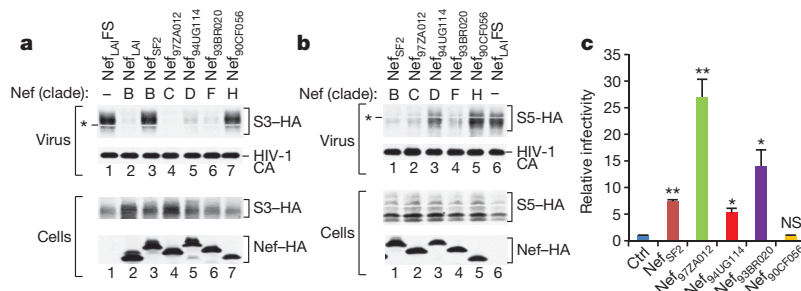


Figure 1 | Inhibition of incorporation of SERINC proteins into HIV-1 virions by Nef correlates with infectivity enhancement. **a, b**, Western blots showing the effects of Nef proteins from various HIV-1 clades on the incorporation of SERINC3–HA (S3–HA) (**a**) or SERINC5–HA (S5–HA) (**b**) into Nef[−] HIV-1 virions. The white bands marked by asterisks are caused by co-migrating HIV-1 Pr55^{gag}. The experiment shown in **a** was performed twice. Supplementary Information contains full scans for **a** and **b**. **c**, Ability of

Nef proteins from different HIV-1 clades to enhance HIV-1 infectivity. Env[−]/Nef[−] HIV-1_{HXB2} particles *trans*-complemented with Env_{HXB2} were produced in JTAG cells in the absence or presence of the indicated Nef proteins, and infectivities normalized for p24 antigen were determined using TZM-bl indicator target cells ($n = 3$). Data are mean and s.d. * $P < 0.05$, ** $P < 0.01$, NS, not significant ($P > 0.05$) (two-tailed unpaired *t*-test with Welch's correction in case of unequal variance; *F*-test, $\alpha = 0.025$). Ctrl, control.

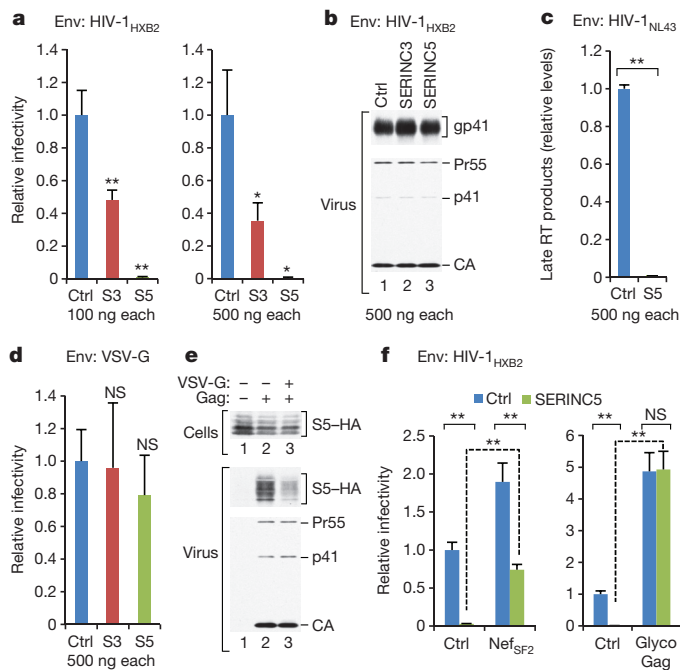


Figure 2 | Effects of exogenous SERINC5 on HIV-1 infectivity.

a, Overexpression of SERINC5 in virus producer cells dramatically reduces Nef⁻ HIV-1 progeny virus single-round infectivity. The effects of exogenous SERINC3 (S3) and SERINC5 (S5) were measured using TZM-bl indicator cells ($n = 3$). **b**, Western blots showing that virus production, Gag processing, and gp41 (Env) incorporation were unaffected. **c**, Nef⁻ HIV-1 progeny virions produced in the presence of exogenous SERINC5 are defective in the synthesis of late reverse transcriptase (RT) products ($n = 2$). **d**, Effects of exogenous SERINC5 on the single-round infectivity of VSV-G-pseudotyped Nef⁻ HIV-1 virions measured as in **a** ($n = 3$). **e**, VSV-G reduces the association of SERINC5-HA with Env⁻ HIV-1 virions. The HIV-1 proviral plasmid in lane 1 has a disrupted *gag* gene. This experiment was repeated twice. Supplementary Information contains full scans for **b** and **e**. **f**, Nef_{SF2} and glycoGag expressed *in trans* in virus producer cells counteract the effect of exogenous SERINC5 on Nef⁻ HIV-1 progeny virion infectivity ($n = 3$). Data are mean and s.d. * $P < 0.05$, ** $P < 0.01$ (two-tailed unpaired *t*-test with Welch's correction in case of unequal variance).

Effects of SERINC depletion on HIV infectivity

JTag T lymphoid cells express both *SERINC3* and *SERINC5* at relatively high levels (Extended Data Fig. 4b). Short interfering RNAs (siRNAs) that knocked down HA-tagged *SERINC3* or *SERINC5* (Fig. 3a) enhanced the specific infectivity of Nef⁻, Env_{HXB2}-pseudotyped HIV-1 particles produced in JTag cells by >4- or >8-fold, respectively (Fig. 3b). In five independent experiments, both siRNAs together enhanced the specific infectivity of Nef⁻ progeny virions 18- to 45-fold (Fig. 3b, d and data not shown). The siRNAs against *SERINC3* and *SERINC5* together also significantly enhanced the infectivity of Nef⁻, Env_{89.6}-bearing HIV-1 virions produced by infected primary macrophages (Fig. 3c). However, they had little effect on the already high specific infectivity of Nef⁻ progeny virions produced in JTag cells that was observed when Nef_{97ZA012} or glycoGag were expressed *in trans* (Fig. 3d).

The Env proteins of the R5-tropic primary HIV-1 isolates SF162 and JRFL differ considerably in their responsiveness to Nef or glycoGag, which is determined by variable regions 1 and 2 (V1/V2) of gp120 (ref. 30). Remarkably, the siRNAs against *SERINC3* and *SERINC5* together precisely phenocopied the differential effects of Nef_{97ZA012} on the specific infectivities of Nef⁻ HIV-1 progeny particles bearing Env_{SF162} or Env_{JRFL} (Fig. 3e). Furthermore, responsiveness to Nef_{97ZA012} and to the siRNAs targeting *SERINC3* and *SERINC5* could be switched simultaneously by exchanging the V1/V2 regions of Env_{SF162} and Env_{JRFL} (Fig. 3e). Nef⁻ HIV-1 virions bearing Env

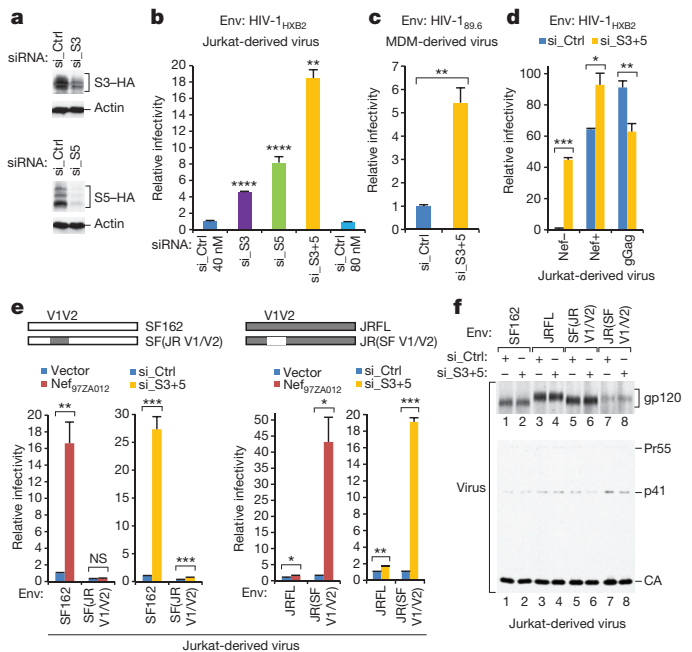


Figure 3 | Effects of depleting SERINC3 and SERINC5 in virus producer cells. **a**, Depletion of HA-tagged SERINC3 and SERINC5 in JTag cells by specific siRNAs. si_Ctrl, non-targeting siRNA control; si_S3, siRNA targeting *SERINC3*; si_S5, siRNA targeting *SERINC5*. **b**, Single-round infectivities of Nef⁻ HIV-1 virions produced in JTag cells ($n = 3$) subjected to non-targeting siRNA or to siRNAs targeting *SERINC3* (si_S3), *SERINC5* (si_S5), or both (si_S3 + S5). **c**, Single-round infectivities of Nef⁻ HIV-1 virions produced in primary monocyte-derived macrophages (MDM) subjected to siRNAs ($n = 3$). **d**, Simultaneous depletion of *SERINC3* and *SERINC5* has negligible effects on Nef⁻ HIV-1 progeny virion infectivity when Nef or glycoGag are provided *in trans* ($n = 3$). **e**, The effects of depleting *SERINC3* together with *SERINC5* on virus infectivity are governed by the same determinants in gp120 V1/V2 that govern Nef-responsiveness ($n = 3$). **f**, Western blots showing that the combined siRNAs targeting *SERINC3* and *SERINC5* did not affect particle production, Gag processing, or Env incorporation. Supplementary Information contains full scans for **a** and **f**. Data are mean and s.d. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$, (two-tailed unpaired *t*-test with Welch's correction in case of unequal variance). The experiments shown in **a** and **b** were performed twice.

proteins that differed profoundly in their responsiveness to Nef or to SERINC depletion incorporated comparable amounts of SERINC5-HA (Extended Data Fig. 7). Furthermore, Nef_{97ZA012} largely prevented the incorporation of SERINC5-HA in the presence of both Env proteins (Extended Data Fig. 7). Importantly, the simultaneous depletion of *SERINC3* and *SERINC5* in virus producer cells affected neither particle morphogenesis nor Env incorporation (Fig. 3f). Collectively, these data indicate that *SERINC3* and *SERINC5* together account for the effects of Nef and glycoGag on HIV-1 infectivity.

HIV infectivity in SERINC knockout cells

Next, we knocked out the *SERINC3* and *SERINC5* genes in JTag cells using the clustered regularly interspaced short palindromic repeats (CRISPR)-Cas9 system³⁹ (Extended Data Fig. 8). Nef⁻, Env_{HXB2}-pseudotyped HIV-1 particles produced in JTag clones lacking either *SERINC3* or *SERINC5* were ~5-fold or 13–20-fold more infectious, respectively, than particles produced in the parental cells (Fig. 4a). Notably, the specific infectivity of particles produced in double-knockout cells lacking *SERINC3* and *SERINC5* was >100-fold higher (Fig. 4a). Furthermore, the markedly increased specific infectivity of viral particles produced in double-knockout cells could be confirmed by visualizing green fluorescence protein (GFP)-positive cells after exposure to recombinant HIV-1 expressing GFP (Fig. 4b). Nef and glycoGag potentially enhanced the specific infectivity of particles produced in parental JTag cells as expected, but had no significant effects

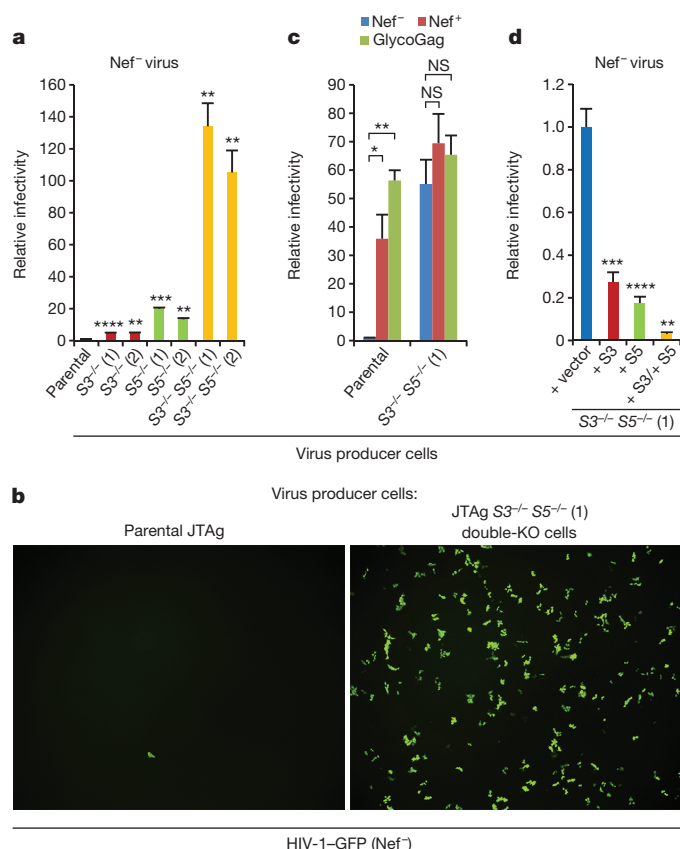


Figure 4 | Effects of *SERINC* knockout and reconstitution on HIV-1 infectivity. **a**, Single-round infectivities of Nef⁻ HIV-1 progeny virions produced in parental or in knockout JTAG cells lacking *SERINC3* (S3^{-/-}), *SERINC5* (S5^{-/-}) or both (S3^{-/-} S5^{-/-}) ($n = 3$). Numbers in parentheses denote clone numbers. **b**, T2M-bl cells were incubated with equal amounts of single-cycle Nef⁻ HIV-1-GFP produced in parental or double-knockout (KO) cells lacking *SERINC3* and *SERINC5*. Infected T2M-bl cells expressing GFP were detected by fluorescence microscopy. **c**, Effects of Nef and glycoGag provided *in trans* on the single-round infectivities of Nef⁻ HIV-1 virions produced in parental or double-knockout cells ($n = 3$). **d**, Effects of introducing expression cassettes for *SERINC3*, *SERINC5* or both into the double-knockout cells on Nef⁻ HIV-1 progeny virus infectivities ($n = 3$). Data are mean and s.d. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$ (two-tailed unpaired *t*-test with Welch's correction in case of unequal variance). The experiments shown in **a** and **b** were repeated three times.

on the already highly infectious particles produced in double-knockout cells (Fig. 4c). The introduction of expression cassettes for *SERINC3*, for *SERINC5*, and for both, into the double-knockout cells via retroviral transduction led to 3.6-fold, 5.7-fold, and 32-fold reductions, respectively, in the specific infectivities of Nef⁻ HIV-1 particles produced in these cells (Fig. 4d). These data confirm that *SERINC3* and *SERINC5* synergistically restrict HIV-1 infectivity in the absence of Nef.

The effects of Nef on HIV-1 replication in cell lines have generally been more modest than in primary lymphocytes^{4,6}. However, apart from Jurkat cells, T cell lines often express only low levels of *SERINC5* mRNA (Extended Data Fig. 4 and data not shown). We observed that at low input virus concentrations, Nef^{NL43} and Nef^{97ZA012} robustly enhanced HIV-1 spreading in highly permissive Jurkat E6.1 cells (Fig. 5a). Nef^{NL43} and Nef^{97ZA012} also enhanced HIV-1 replication in JTAG cells, as judged from Gag protein expression levels in the infected cells and from the release of p24 antigen over time (Fig. 5b, c). In marked contrast, the Nef⁺ and Nef⁻ viruses replicated with similar kinetics in double-knockout JTAG cells lacking *SERINC3* and *SERINC5*, which were generally more permissive than the parental cells (Fig. 5b, c). Crucially, the requirement for Nef was restored in

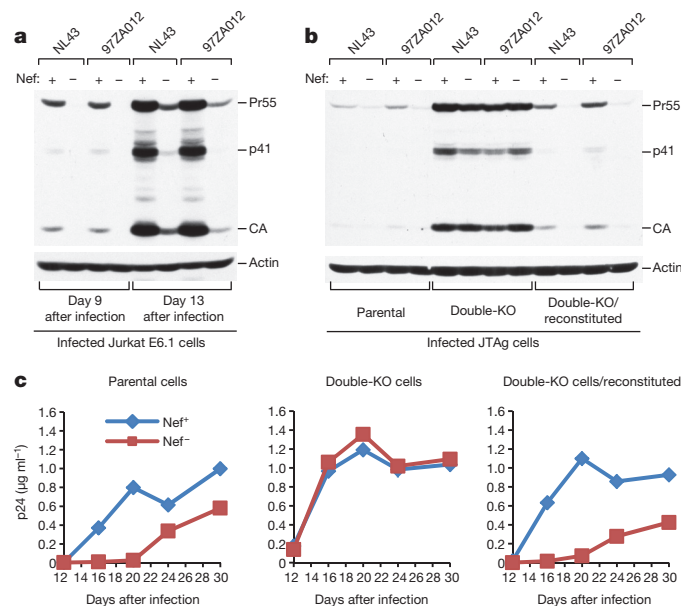


Figure 5 | Nef counteracts inhibition of HIV-1 replication by *SERINC3* and *SERINC5*. **a**, Effect of Nef on HIV-1 spreading in Jurkat E6.1 cells infected at a low input virus concentration (100 pg ml⁻¹ p24). Gag protein expression in the cultures 9 and 13 days after infection was examined by western blotting as a measure of virus replication. **b**, **c**, Effects of Nef on virus spreading in parental JTAG cells, double-knockout cells lacking *SERINC3* and *SERINC5*, and *SERINC3*+*SERINC5*-reconstituted double-knockout cells. The spreading of HIV-1_{NL43}-based viruses encoding either wild-type or disrupted versions of Nef_{NL43} or Nef_{97ZA012} was examined by western blotting of cell lysates with anti-CA antibody 9 days after infection with 20 ng ml⁻¹ p24 (**b**), or by monitoring p24 accumulation in the supernatants after infection with 4 ng ml⁻¹ p24 (**c**). Relatively high input virus concentrations were used to compensate for low CD4 levels on JTAG cells. The experiment shown in **b** was repeated twice. Supplementary Information contains full scans for **a** and **b**.

double-knockout cells reconstituted with *SERINC3* and *SERINC5* expression cassettes (Fig. 5b, c). While the levels of *SERINC3* and *SERINC5* in the reconstituted cells were higher than in the parental cells (Extended Data Fig. 9), they were comparable to those in human PBMC (Extended Data Fig. 4).

Although endogenous CD4 levels in JTAG cells are low, similar results were obtained with more permissive CD4^{high} versions generated by retroviral transduction. In the presence of extra CD4, Nef again clearly enhanced virus replication in parental JTAG cells, but was entirely dispensable in double-knockout cells lacking *SERINC3* and *SERINC5* (Extended Data Fig. 10). Furthermore, the role of Nef in virus replication was restored after reconstitution of *SERINC3* and *SERINC5* expression in the double-knockout cells (Extended Data Fig. 10). These results demonstrate that *SERINC3* and *SERINC5* together restrict HIV-1 replication, and that Nef antagonizes this restriction.

Discussion

Our findings reveal that HIV-1 Nef and MLV glycoGag efficiently downregulate *SERINC3* and *SERINC5* from the cell surface, which prevents their incorporation into HIV-1 virions and consequently counteracts their inhibitory effect on HIV-1 infectivity. Importantly, these findings offer an explanation for why the enhancement of HIV-1 infectivity by Nef and glycoGag is highly dependent on dynamin 2, clathrin and the AP-2 clathrin adaptor complex^{29,31}. *SERINC* family members are present in all eukaryotes, but their functions remain largely unknown. *SERINC* proteins reportedly enhance the incorporation of serine into phosphatidylserine and sphingolipids³³. In principle, this activity could affect the lipid composition of the viral envelope, which is considered crucial for virion infectivity⁴⁰. Our data demonstrate that

SERINC3 and SERINC5 together account for most if not all of the effects of Nef on HIV-1 infectivity and on HIV-1 replication in JTag cells. Notably, Nef enhances HIV-1 infectivity and stimulates HIV-1 replication in human PBMC^{4–6,18,41}, whose *SERINC3* and *SERINC5* mRNA levels exceed those of Jurkat cells (Extended Data Fig. 4).

The ability of virions produced in the absence of Nef to reverse transcribe their genome in target cells is impaired^{17,19,20}. Consistent with these observations, we find that *SERINC5* in virus producer cells strongly inhibits the ability of Nef[−] HIV-1 virions to complete reverse transcription. We also find that *SERINC5* can in principle abolish the ability of progeny HIV-1 virions to fuse with target cells. However, lower levels of *SERINC5* inhibited the fusion step to a lesser extent than the ability of progeny virions to productively infect target cells. Although there is controversy about the effect of Nef on HIV-1 entry^{35,42,43}, a twofold reduction in the ability of Nef[−] HIV-1 virions to fuse with target cells was noted in one study³⁵. In all of these studies, virus was produced in 293T cells, whose endogenous *SERINC5* mRNA levels are low (Extended Data Fig. 4). It is conceivable that relatively low levels of virion-associated *SERINC5* impair primarily fusion pore enlargement, which poses a higher energy barrier to overcome than pore formation⁴⁴. This would be expected to impair passage of the viral core but not necessarily of the much smaller BlaM-Vpr fusion indicator into target cells. Consistent with a role in entry, Nef enhances the cytoplasmic delivery of viral cores⁴⁵. Further, the requirement for Nef is determined by HIV-1 Env³⁰.

Interestingly, the Env proteins of HIV-1_{NL43} and HIV-1_{SF162}, which are highly responsive to Nef and glycoGag, require a higher number of Env trimers to complete entry than the poorly Nef/glycoGag-responsive Env_{JRFL}^{30,46}. Furthermore, the naturally occurring Asn160Lys mutation in the V2 loop of HIV-1 Env, which results in the loss of a glycosylation site, can increase both the responsiveness to Nef and glycoGag and the stoichiometry of entry^{30,46}. Mechanistically, a link between Nef/glycoGag responsiveness and the stoichiometry of entry could be due to an inhibitory effect of virion-associated SERINC3 on the clustering of Env trimers. Notably, such clusters have been visualized on the surface of mature HIV-1 virions and, most prominently, at virus-cell contact zones^{47,48}. Alternatively, SERINC3 embedded in the virion membrane could increase the energy barrier for fusion pore expansion. In both scenarios, differences in Nef/glycoGag-responsiveness among HIV-1 Envs, as well as differences in SERINC-sensitivity, may ultimately be due to differences in the amount of energy that these Envs provide towards fusion. Regardless of the mechanism, our observation that viruses as distant as HIV-1 and MLV have evolved to counteract *SERINC3* and *SERINC5* raises the possibility that these proteins have a broader role in innate antiviral immunity.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 28 May; accepted 18 August 2015.

Published online 30 September 2015.

- Kestier, H. W. III *et al.* Importance of the *nef* gene for maintenance of high virus loads and for development of AIDS. *Cell* **65**, 651–662 (1991).
- Deacon, N. J. *et al.* Genomic structure of an attenuated quasi species of HIV-1 from a blood transfusion donor and recipients. *Science* **270**, 988–991 (1995).
- Zou, W. *et al.* Nef functions in BLT mice to enhance HIV-1 replication and deplete CD4⁺CD8⁺ thymocytes. *Retrovirology* **9**, 44 (2012).
- Kim, S., Ikeuchi, K., Byrn, R., Groopman, J. & Baltimore, D. Lack of a negative influence on viral growth by the *nef* gene of human immunodeficiency virus type 1. *Proc. Natl Acad. Sci. USA* **86**, 9544–9548 (1989).
- Miller, M. D., Warmerdam, M. T., Gaston, I., Greene, W. C. & Feinberg, M. B. The human immunodeficiency virus-1 *nef* gene product: a positive factor for viral infection and replication in primary lymphocytes and macrophages. *J. Exp. Med.* **179**, 101–113 (1994).
- Spina, C. A., Kwah, T. J., Chow, M. Y., Guatelli, J. C. & Richman, D. D. The importance of *nef* in the induction of human immunodeficiency virus type 1 replication from primary quiescent CD4 lymphocytes. *J. Exp. Med.* **179**, 115–123 (1994).
- Garcia, J. V. & Miller, A. D. Serine phosphorylation-independent downregulation of cell-surface CD4 by *nef*. *Nature* **350**, 508–511 (1991).

- Aiken, C., Konner, J., Landau, N. R., Lenburg, M. E. & Trono, D. Nef induces CD4 endocytosis: requirement for a critical dileucine motif in the membrane-proximal CD4 cytoplasmic domain. *Cell* **76**, 853–864 (1994).
- Rhee, S. S. & Marsh, J. W. Human immunodeficiency virus type 1 Nef-induced down-modulation of CD4 is due to rapid internalization and degradation of surface CD4. *J. Virol.* **68**, 5156–5163 (1994).
- Chaudhuri, R., Lindwasser, O. W., Smith, W. J., Hurley, J. H. & Bonifacio, J. S. Downregulation of CD4 by human immunodeficiency virus type 1 Nef is dependent on clathrin and involves direct interaction of Nef with the AP2 clathrin adaptor. *J. Virol.* **81**, 3877–3890 (2007).
- Veillette, M. *et al.* Interaction with cellular CD4 exposes HIV-1 envelope epitopes targeted by antibody-dependent cell-mediated cytotoxicity. *J. Virol.* **88**, 2633–2644 (2014).
- Haller, C. *et al.* HIV-1 Nef and Vpu are functionally redundant broad-spectrum modulators of cell surface receptors, including tetraspanins. *J. Virol.* **88**, 14241–14257 (2014).
- Schwartz, O., Marechal, V., Le Gall, S., Lemonnier, F. & Heard, J. M. Endocytosis of major histocompatibility complex class I molecules is induced by the HIV-1 Nef protein. *Nature Med.* **2**, 338–342 (1996).
- Collins, K. L., Chen, B. K., Kalams, S. A., Walker, B. D. & Baltimore, D. HIV-1 Nef protein protects infected primary cells against killing by cytotoxic T lymphocytes. *Nature* **391**, 397–401 (1998).
- Cohen, G. B. *et al.* The selective downregulation of class I major histocompatibility complex proteins by HIV-1 protects HIV-infected cells from NK cells. *Immunity* **10**, 661–671 (1999).
- Schindler, M. *et al.* Nef-mediated suppression of T cell activation was lost in a lentiviral lineage that gave rise to HIV-1. *Cell* **125**, 1055–1067 (2006).
- Chowers, M. Y., Pandori, M. W., Spina, C. A., Richman, D. D. & Guatelli, J. C. The growth advantage conferred by HIV-1 *nef* is determined at the level of viral DNA formation and is independent of CD4 downregulation. *Virology* **212**, 451–457 (1995).
- Münch, J. *et al.* Nef-mediated enhancement of virion infectivity and stimulation of viral replication are fundamental properties of primate lentiviruses. *J. Virol.* **81**, 13852–13864 (2007).
- Aiken, C. & Trono, D. Nef stimulates human immunodeficiency virus type 1 proviral DNA synthesis. *J. Virol.* **69**, 5048–5056 (1995).
- Schwartz, O., Marechal, V., Danos, O. & Heard, J. M. Human immunodeficiency virus type 1 Nef increases the efficiency of reverse transcription in the infected cell. *J. Virol.* **69**, 4053–4059 (1995).
- Miller, M. D., Warmerdam, M. T., Page, K. A., Feinberg, M. B. & Greene, W. C. Expression of the human immunodeficiency virus type 1 (HIV-1) *nef* gene during HIV-1 production increases progeny particle infectivity independently of gp160 or viral entry. *J. Virol.* **69**, 579–584 (1995).
- Forshey, B. M. & Aiken, C. Disassembly of human immunodeficiency virus type 1 cores *in vitro* reveals association of Nef with the subviral ribonucleoprotein complex. *J. Virol.* **77**, 4409–4414 (2003).
- Ross, T. M., Oran, A. E. & Cullen, B. R. Inhibition of HIV-1 progeny virion release by cell-surface CD4 is relieved by expression of the viral Nef protein. *Curr. Biol.* **9**, 613–621 (1999).
- Lama, J., Mangasarian, A. & Trono, D. Cell-surface expression of CD4 reduces HIV-1 infectivity by blocking Env incorporation in a Nef- and Vpu-inhibitable manner. *Curr. Biol.* **9**, 622–631 (1999).
- Goldsmith, M. A., Warmerdam, M. T., Atchison, R. E., Miller, M. D. & Greene, W. C. Dissociation of the CD4 downregulation and viral infectivity enhancement functions of human immunodeficiency virus type 1 Nef. *J. Virol.* **69**, 4112–4121 (1995).
- Pizzato, M. MLV glycosylated-Gag is an infectivity factor that rescues Nef-deficient HIV-1. *Proc. Natl Acad. Sci. USA* **107**, 9364–9369 (2010).
- Prats, A. C., De Billy, G., Wang, P. & Darlix, J. L. CUG initiation codon used for the synthesis of a cell surface antigen coded by the murine leukemia virus. *J. Mol. Biol.* **205**, 363–372 (1989).
- Pillemer, E. A., Kooistra, D. A., Witte, O. N. & Weissman, I. L. Monoclonal antibody to the amino-terminal L sequence of murine leukemia virus glycosylated gag polyproteins demonstrates their unusual orientation in the cell membrane. *J. Virol.* **57**, 413–421 (1986).
- Usami, Y., Popov, S. & Gottlinger, H. G. The Nef-like effect of murine leukemia virus glycosylated gag on HIV-1 infectivity is mediated by its cytoplasmic domain and depends on the AP-2 adaptor complex. *J. Virol.* **88**, 3443–3454 (2014).
- Usami, Y. & Gottlinger, H. HIV-1 Nef Responsiveness Is Determined by Env Variable Regions Involved in Trimer Association and Correlates with Neutralization Sensitivity. *Cell Rep.* **5**, 802–812 (2013).
- Pizzato, M. *et al.* Dynamin 2 is required for the enhancement of HIV-1 infectivity by Nef. *Proc. Natl Acad. Sci. USA* **104**, 6812–6817 (2007).
- Craig, H. M., Pandori, M. W. & Guatelli, J. C. Interaction of HIV-1 Nef with the cellular dileucine-based sorting pathway is required for CD4 down-regulation and optimal viral infectivity. *Proc. Natl Acad. Sci. USA* **95**, 11229–11234 (1998).
- Inuzuka, M., Hayakawa, M. & Ingi, T. Serinc, an activity-regulated protein family, incorporates serine into membrane lipid synthesis. *J. Biol. Chem.* **280**, 35776–35783 (2005).
- Grossman, T. R., Luque, J. M. & Nelson, N. Identification of a ubiquitous family of membrane proteins and their expression in mouse brain. *J. Exp. Biol.* **203**, 447–457 (2000).
- Day, J. R., Munk, C. & Guatelli, J. C. The membrane-proximal tyrosine-based sorting signal of human immunodeficiency virus type 1 gp41 is required for optimal viral infectivity. *J. Virol.* **78**, 1069–1079 (2004).

36. Cavois, M., De Noronha, C. & Greene, W. C. A sensitive and specific enzyme-based assay detecting HIV-1 virion fusion in primary T lymphocytes. *Nature Biotechnol.* **20**, 1151–1154 (2002).
37. Aiken, C. Pseudotyping human immunodeficiency virus type 1 (HIV-1) by the glycoprotein of vesicular stomatitis virus targets HIV-1 entry to an endocytic pathway and suppresses both the requirement for Nef and the sensitivity to cyclosporin A. *J. Virol.* **71**, 5871–5877 (1997).
38. Luo, T., Douglas, J. L., Livingston, R. L. & Garcia, J. V. Infectivity enhancement by HIV-1 Nef is dependent on the pathway of virus entry: implications for HIV-based gene transfer systems. *Virology* **241**, 224–233 (1998).
39. Mali, P., Esvelt, K. M. & Church, G. M. Cas9 as a versatile tool for engineering biology. *Nature Methods* **10**, 957–963 (2013).
40. Waheed, A. A. & Freed, E. O. The Role of Lipids in Retrovirus Replication. *Viruses* **2**, 1146–1180 (2010).
41. Lundquist, C. A., Tobiume, M., Zhou, J., Unutmaz, D. & Aiken, C. Nef-mediated downregulation of CD4 enhances human immunodeficiency virus type 1 replication in primary T lymphocytes. *J. Virol.* **76**, 4625–4633 (2002).
42. Tobiume, M., Lineberger, J. E., Lundquist, C. A., Miller, M. D. & Aiken, C. Nef does not affect the efficiency of human immunodeficiency virus type 1 fusion with target cells. *J. Virol.* **77**, 10645–10650 (2003).
43. Cavois, M., Neideman, J., Yonemoto, W., Fenard, D. & Greene, W. C. HIV-1 virion fusion assay: uncoating not required and no effect of Nef on fusion. *Virology* **328**, 36–44 (2004).
44. Cohen, F. S. & Melikyan, G. B. The energetics of membrane fusion from binding, through hemifusion, pore formation, and pore enlargement. *J. Membr. Biol.* **199**, 1–14 (2004).
45. Schaeffer, E., Geleziunas, R. & Greene, W. C. Human immunodeficiency virus type 1 Nef functions at the level of virus entry by enhancing cytoplasmic delivery of virions. *J. Virol.* **75**, 2993–3000 (2001).
46. Brandenburg, O. F., Magnus, C., Rusert, P., Regoes, R. R. & Trkola, A. Different infectivity of HIV-1 strains is linked to number of envelope trimers required for entry. *PLoS Pathog.* **11**, e1004595 (2015).
47. Sougrat, R. *et al.* Electron tomography of the contact between T cells and SIV/HIV-1: implications for viral entry. *PLoS Pathog.* **3**, e63 (2007).
48. Chojnacki, J. *et al.* Maturation-dependent HIV-1 surface protein redistribution revealed by fluorescence nanoscopy. *Science* **338**, 524–528 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank J. Leszyk and S. Shaffer for protein microsequencing, BGI Americas for RNA-seq, R. Maehr for sgRNA and Cas9 expression plasmids, J. Sodroski for HIVec2.GFP, T. Akagi for pCXbsr, and the AIDS Research and Reference Reagent Program, Division of AIDS, NIAID, NIH for p89.6, indinavir, maraviroc, the monoclonal antibodies 183-H12-5C and Chessie 8, and TZM-bl cells. This work was supported by NIAID/NIH grant R01AI029873 and by NIDA/NIH grant DP1DA038034.

Author Contributions Y.U., Y.W. and H.G.G. designed the experiments and analysed the data. Y.U. carried out the analysis of virions and the SERINC overexpression and depletion experiments. Y.W. generated and characterized the SERINC knockout cells, carried out all experiments involving knockout cells and primary cells, and performed the qRT-PCR experiments and the BlaM-Vpr-based fusion assays. H.G.G. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to H.G.G. (heinrich.gottlinger@umassmed.edu).

METHODS

No statistical methods were used to predetermine sample size. Investigators were not blinded to allocation during experiments and outcome assessment, and experiments were not randomized.

Cells. JTAG, 293T, MT4, A549 and U2-OS cells were gifts from G. Crabtree, D. Baltimore, W. Haseltine, M. Bujny and A. Brass, respectively. Jurkat E6.1 and MOLT-3 cells were obtained from the ATCC. TZM-bl cells were obtained from the AIDS Research and Reference Reagent Program, Division of AIDS, NIAID, NIH. PBMC were isolated from the blood of healthy donors by Ficoll-Hypaque density gradient centrifugation. The MycoSensor PCR assay system (Agilent) was used to check all cell lines for mycoplasma. The cell lines were not authenticated for this study.

HIV-1 proviral constructs. NL4-3/Nefstop, the *nef*-deficient variant of the prototypic HIV-1_{NL4-3} used in this study, has *nef* codons 31–33 replaced by three consecutive premature termination codons. NL-nef_{97ZA012} is a version of HIV-1_{NL4-3} that has the *nef* gene precisely replaced by that of p97ZA012.1 (GenBank accession AF286227), a near full-length molecular clone of a primary subtype C HIV-1 isolate⁴⁹. NL-nef_{97ZA012}FS is a *nef*-deficient variant of NL-nef_{97ZA012} owing to a frameshift at a unique XhoI site in *nef*. HXB/89.6_{ecto}-GFP is a macrophage-tropic variant of the infectious HIV-1 molecular clone HXBH10 that encodes GFP within the *nef* region, and that has a KpnI-BamHI fragment (nucleotides 6351–8475 of K03455) encoding the Env ectodomain replaced by the corresponding fragment from p89.6, a biologically active molecular clone of the primary HIV-1_{89.6} isolate⁵⁰. The proviral plasmids NL4-3/glycoMA, HXB/Env⁺/Nef⁺, HXB/Env⁺/Nef⁺, and HXBH10-gag⁺ have been described^{29,51,52}.

Expression plasmids. The pBJ5-based Nef expression vectors used, the *nef*-deficient control vector pNef_{Δ1}FS, and the pBJ5-based expression vectors for glycoGag–HA, for wild-type glycoMA–HA, and for its truncation mutants have been described^{29,31}. The latter plasmids were used as templates to amplify fragments encoding carboxy-terminally Flag-tagged versions of wild-type glycoMA and of its mutants, which were also cloned into the mammalian expression vector pBJ5. The HIV-1 Env expression vectors used have also been described³⁰. The coding sequences for *SERINC3* and *SERINC5* without or with a C-terminal HA-tag were amplified from BC006088 (*SERINC3*) and from BC101281 and AW005635 (*SERINC5*) (GE Healthcare). The primers included a Kozak sequence and XhoI and NotI cloning sites for insertion into pBJ5. The vectors expressing *SERINC5*(iHA) and *SERINC5*–mCherry are also pBJ5-based. *SERINC5*(iHA) has an HA tag inserted between residues 290 and 291 of *SERINC5*. *SERINC5*–mCherry has a Thr–Gly–Ala–Gly linker inserted between *SERINC5* and mCherry.

Retroviral vectors. The human *SERINC3* and *SERINC5* coding sequences preceded by a Kozak sequence were inserted into pMSCVhyg and pMSCVpuro, respectively (Clontech). The human *CD4* coding sequence was inserted into the retroviral vector pCXbsr⁵³.

Protein identification. For the identification of virus-associated host proteins, virions released by chronically infected T-lymphoid MOLT-3 cells were pelleted through sucrose, resuspended in PBS, and further purified in OptiPrep velocity gradients as described⁵⁴. OptiPrep gradient fractions were collected from the top and diluted with PBS. Viral particles were harvested from the fractions by ultracentrifugation and lysed in SDS–PAGE loading buffer (60 mM Tris–HCl, pH 6.8, 1% SDS, 10% (v/v) glycerol, 0.005% bromophenol blue, 5% (v/v) 2-mercaptoethanol). Virus-containing fractions were then identified by western blotting with antibody 183–H12–5C against HIV-1 capsid (CA)⁵⁵.

For mass spectrometry, virion-associated proteins were briefly run into an SDS–PAGE gel to allow removal of SDS. After in-gel digestion with trypsin, peptides were separated on a NanoAcquity (Waters) UPLC and analysed with a Q Exactive hybrid mass spectrometer (Thermo). The run conditions followed the ‘sensitive’ settings recommended for optimizing the Q Exactive for low abundance proteins⁵⁶. Raw data files were peak processed with Proteome Discoverer (version 1.3, Thermo) before searching with Mascot Server (version 2.4) against the SwissProt database. Search results were then loaded into the Scaffold Viewer (Proteome Software, Inc.).

Viral particle analysis. To examine the incorporation of *SERINC*s, 293T cells were co-transfected with NL4-3/Nefstop, vectors expressing HA-tagged *SERINC*s, and vectors expressing various epitope-tagged Nef or glycoGag proteins, or the appropriate control vectors. To examine whether VSV-G affects *SERINC5* incorporation, 293T cells were co-transfected with 1 μg HXBH10-gag⁺ (a control HIV-1 proviral construct unable to express Gag) or HXB/Env⁺/Nef⁺, 100 ng of a plasmid expressing VSV-G or control vector, and 500 ng of a plasmid expressing *SERINC5*–HA. Virions released into the medium were pelleted through sucrose, and virus- and cell-associated proteins were detected by western blotting as described previously⁵⁷. Samples used for the detection of *SERINC*s were maximally heated to 37 °C, because *SERINC* proteins are highly aggregation-prone

at higher temperatures (data not shown). In some cases, 25 mM TCEP was used as the reducing agent. The antibodies used were 183–H12–5C against HIV-1 CA, HA.11 (Covance) against the HA epitope, M2 against the Flag epitope (Sigma-Aldrich), and AC-40 (Sigma-Aldrich) against actin. To examine Env incorporation, virions produced by transiently transfected 293T or JTAG cells were pelleted through 20% sucrose cushions by ultracentrifugation, and examined by western blotting using the anti-gp41 monoclonal antibody Chessie 8 (ref. 58), an anti-gp120 polyclonal antibody (20–HG81; Fitzgerald), and the anti-CA monoclonal antibody 183–H12–5C.

SERINC overexpression experiments. Pseudovirions capable of a single round of replication were produced by transfecting 293T cells in triplicate using a calcium phosphate precipitation method. The cells were co-transfected with 1 μg HXB/Env⁺/Nef⁺, 100 ng of a plasmid expressing Env_{HXB2} or VSV-G, and 100 or 500 ng of plasmids expressing *SERINC3* or *SERINC5*, or with equimolar amounts of the empty vector. To examine the effects of Nef or glycoGag, 293T cells were co-transfected with 1 μg HXB/Env⁺/Nef⁺, a plasmid expressing Env_{HXB2} (100 ng), a plasmid expressing *SERINC5* (100 ng) or the empty vector, and plasmids expressing Nef_{ΔF2} (2 μg) or glycoGag (200 ng) or the empty vector. Supernatants containing progeny virions were collected two days after transfection, clarified by low-speed centrifugation, filtered through 0.45-μm pore filters, and then used immediately to infect TZM-bl indicator cells in T25 flasks. Aliquots of the filtered virus stocks were frozen for HIV-1 CA (p24) antigen quantitation by a standard ELISA. Three to five days after infection, the indicator cells were lysed in reporter lysis buffer (Promega), and β-galactosidase activity was determined as a measure of infection using a kit (E2000; Promega) according to the manufacturer's instructions.

To examine the effects of exogenous *SERINC5* on the single-cycle infectivity of *nef*-deficient HIV-1 for primary target cells, viral stocks were obtained by co-transfecting 293T cells with HXB/Env⁺/Nef⁺, a plasmid expressing Env_{HXB2}, a plasmid expressing *SERINC5* or the empty vector, and an HIV-1-based lentiviral vector expressing GFP. Filtered viral stocks normalized for p24 antigen were used to infect human PBMC, and infected cells expressing GFP were quantified by flow cytometry.

SERINC depletion experiments. To obtain pseudovirions capable of a single round of replication, Lipofectamine 2000 (Invitrogen) was used to transfect JTAG cells in triplicate with 1 μg HXB/Env⁺/Nef⁺, 100 ng of an HIV-1 Env expression plasmid, and siRNAs (40 nM each). Additionally, 500 ng of a plasmid expressing Nef_{97ZA012}, or 200 ng of a plasmid expressing glycoGag, or the empty pBJ5 expression plasmid, were co-transfected in some experiments. The siRNAs targeting *SERINC3* (Hs_TDE1_2; target sequence: 5'-CACGGTGACTCGCCT CATTTA-3') or *SERINC5* (Hs_C5orf12_3; target sequence: 5'-CACGCTCT ACATCTACTCCTA-3'), and AllStars negative control siRNA were purchased from Qiagen. As a control for experiments in which the siRNAs targeting *SERINC3* and *SERINC5* were co-transfected, the concentration of the control siRNA was doubled. The infectivities of JTAG-derived virus stocks normalized for p24 antigen content were determined as above using TZM-bl indicator cells.

To examine the effects of *SERINC* proteins on the infectivity of HIV-1 progeny virions produced in primary cells, monocyte-derived macrophages were infected with replication-competent, dual-tropic HXB/89.6_{ecto}-GFP. On day 5 after infection, Lipofectamine 2000 was used to simultaneously transfect the monocyte-derived macrophages with the siRNAs targeting *SERINC3* and *SERINC5* (240 nM each), or with the negative control siRNA. The cells were washed 5 h later to remove the transfection agent. Virus-containing culture medium was collected on day 3 after transfection, and infectivities normalized for p24 antigen were determined using TZM-bl indicator cells. Indinavir (2 μM) was added to the TZM-bl cells together with virus to limit replication to a single cycle, and AMD3100 (5 μM) and maraviroc (50 nM) were added the next day to prevent Env-induced cell–cell fusion.

Generation and use of knockout cells. Expression plasmids for single-guide RNAs (sgRNAs) targeting exons within the *SERINC3* and *SERINC5* genes were transiently transfected into JTAG cells by nucleofection, along with a plasmid expressing Cas9. The sites targeted by the sgRNAs are depicted in Extended Data Fig. 8. Whereas the two sgRNAs targeting the *SERINC3* gene were expressed individually, the two sgRNAs targeting the *SERINC5* gene were expressed together. To obtain double-knockout cells, JTAG S3^{−/−} (2) cells were co-transfected with the two sgRNAs targeting the *SERINC5* gene and the Cas9 expression plasmid. Nine days after transfection, gene editing in the bulk cultures was confirmed by PCR amplification of the targeted regions of the genome, followed by digestion of the PCR products with appropriate restriction enzymes (NcoI and BtsCI for target sites A and B within the *SERINC3* gene, respectively; BsoBI for target site B within the *SERINC5* gene). Clones were then obtained by limiting dilution in 96-well flat-bottomed culture plates. Whenever possible, the clones were pre-screened by PCR amplification of the targeted

regions of the genome and restriction analysis. Furthermore, the PCR products were cloned into pCR-Blunt II-TOPO (Invitrogen/Life Technologies), and up to 10 independent clones were sequenced in each case. The primer pairs used for PCR amplification of the sgRNA target sites were: 5'-CCATAGTCAGTCTTG CAGTTG-3' and 5'-GTACGTAGTATCTAGCATAGTGC-3' (*SERINC3* target site A), 5'-CTTCTAGGCTAATGTTGTCC-3' and 5'-GTGAGTTGCAGGTA CTAAGTC-3' (*SERINC3* target site B), 5'-CACACGATCCATTTCCACAG-3' and 5'-CGCATCATGGTACCAGGTG-3' (*SERINC5* target site A), and 5'-GATCATTGGCAGGTAAGAGC-3' and 5'-CACACGCAAACACAAGC-3' (*SERINC5* target site B). Deletions between *SERINC5* target sites A and B were identified using primers 5'-CACACGATCCATTTCCACAG-3' and 5'-CACAC CGCAAACACAAGC-3' for PCR amplification and direct sequencing of the products. An inversion between *SERINC5* target sites A and B was characterized using primer pair 5'-CACACGATCCATTTCCACAG-3' and 5'-GATCATTGG CAGGTAAGAGC-3', and primer pair 5'-CGCATCATGGTACCAGGTG-3' and 5'-CACACGCAAACACAAGC-3'. Ectopic *SERINC* expression cassettes were introduced into the double-knockout cells by retroviral transduction with MSCVhyg*SERINC3* and/or MSCVpuro*SERINC5*, followed by selection with hygromycin and/or puromycin. *SERINC3* expression was examined by western blotting with a rabbit anti-TDE1 (*SERINC3*) antibody (GTX115512; GeneTex).

To determine the effects of *SERINC* proteins on HIV-1 infectivity in the absence of Nef, parental, knockout, double-knockout, and gene-reconstituted double-knockout JTAG cells were co-transfected in triplicate with a plasmid expressing Env_{HXB2} and HXB/Env⁻/Nef⁻. To determine the effects of Nef and glycoGag in cells lacking *SERINC* genes, parental and double-knockout JTAG cells were co-transfected with a plasmid expressing Env_{HXB2} and HXB/Env⁻/Nef⁻, HXB/Env⁻/Nef⁺ or HXB/Env⁻/Nef⁻ together with a plasmid expressing glycoGag. Progeny virus infectivities normalized for p24 antigen were determined using TZM-bl indicator cells. Alternatively, the HIV-1 vector HIVec2.GFP was co-transfected together with HXB/Env⁻/Nef⁻ and the Env expression plasmid. After exposure to equal amounts of virus, infected TZM-bl cells were then identified based on GFP expression.

For virus replication studies, replication-competent HIV-1 was produced by transiently transfecting 293T cells with NL4-3, NL4-3/Nefstop, NL-nef_{97ZA012} or NL-nef_{97ZA012}FS. Virus-containing supernatants were passed through 0.45-µm filters, normalized for p24 antigen, and used to infect parental, double-knockout and gene-reconstituted double-knockout JTAG cells, or CD4^{high} versions obtained by retroviral transduction with pCXbsrCD4 and selection with blasticidin.

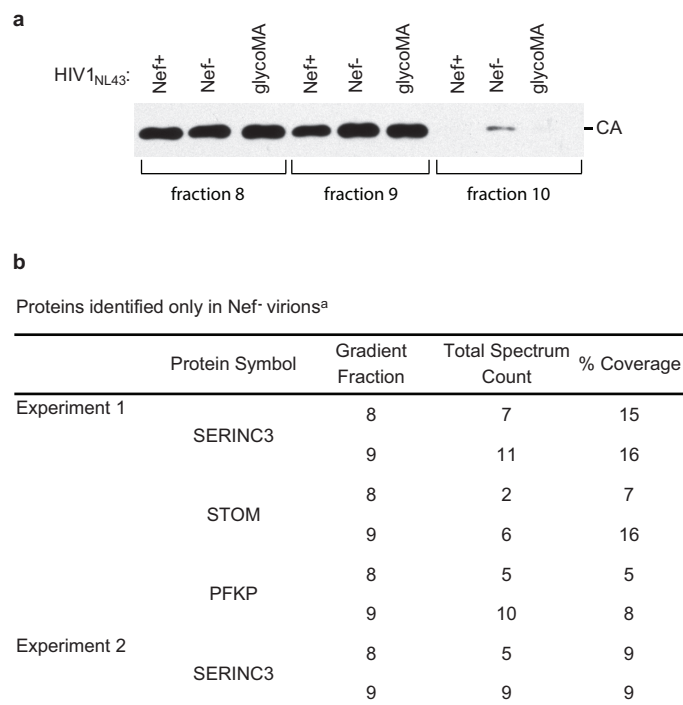
Analysis of mRNA expression. Total cellular RNA was extracted from cell lines and PBMC using an RNeasy mini kit (Qiagen) and treated with RNase-free DNase (Qiagen). The A_{260 nm}/A_{280 nm} ratio was >2.0 for all samples analysed. Quantitative reverse transcription PCR (qRT-PCR) was performed in triplicate for each biological sample using a LightCycler 96 real-time PCR system (Roche) and a Kapa SYBR FAST One-Step qRT-PCR Universal kit (Kapa Biosystems) according to the manufacturer's instructions. Threshold cycle values were normalized for those obtained for *GAPDH*, and relative expression levels were calculated using the 2^{-ΔΔC_T} method⁵⁹. The primer pairs used were: *SERINC3*, 5'-AATTCAGGAACACCAGCCTC-3' and 5'-GGTTGGGATTGCAGGAAC GA-3'; *SERINC5*, 5'-ATCGAGTTCTGACGCTCTGC-3' and 5'-GCTCTTC AGTGCTCTCTCCAC-3'; *GAPDH*, 5'-TGCACCACCAACTGCTTAGC-3' and 5'-GGCATGGACTGTGGTCATGAG-3'.

Analysis of late reverse transcriptase products. Virions were produced by co-transfecting 293T cells with NL4-3/Nefstop (1.5 µg) and the pBJ5-based vector expressing *SERINC5* (500 ng) or an equimolar amount of empty pBJ5. Cell-free virions were treated with RNase-free DNase I (Roche), and used to infect A549/CD4/CXCR4 cells in duplicate in T25 flasks for 14 h in the absence or presence of a cocktail of reverse transcriptase inhibitors. Genomic DNA was extracted with DNAzol (Life Technologies), and 100 ng of each template DNA was used for quantitative PCR using a LightCycler 96 real-time qPCR system (Roche) and a

Kapa SYBR FAST Universal qPCR kit (Kapa Biosystems) according to the manufacturer's instructions. The primers used to quantify late reverse transcriptase products were J1 forward 5'-ACAAGCTAGTACCAGTTGAGCCAGATAAG-3', and J2 reverse 5'-GCCGTGCGCGCTTCAGCAAGC-3'. The J1 forward primer exploits differences between the 5' and 3' long terminal repeats of pNL4-3 to help distinguish between late reverse transcriptase products and contaminating plasmid DNA. Standard curves were obtained from tenfold serial dilutions of DNA extracted from cells infected with virions produced in the absence of exogenous *SERINC5*. Quantitative PCR results were normalized for input virus based on p24 antigen quantifications. A549/CD4/CXCR4 target cells were generated by transduction with retroviral vectors expressing CD4 (pMSCVpuroCD4) and CXCR4 (pCXbsrCXCR4).

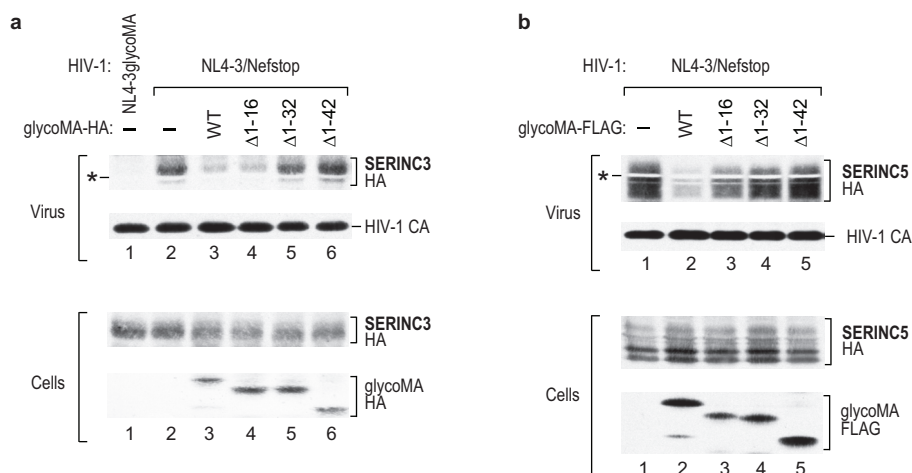
Virion fusion assay. Virions containing BlaM-Vpr were produced by transfecting 293T cells with HXB/Env⁻/Nef⁻ (2.5 µg), a vector expressing the Env protein of HIV-1_{HXB2} or a frameshifted version unable to express Env (200 ng), the BlaM-Vpr expression vector pMM310 (1 µg), and a pBJ5-based vector expressing *SERINC5* (1 µg or 100 ng) or an equimolar amount of empty pBJ5 (0.7 µg or 70 ng). Cell-free virions were normalized for p24 antigen and incubated with 2 × 10⁵ TZM-bl or A549/CD4/CXCR4 cells in 6-well plates for 4 h at 37 °C. After washing with PBS, 1 ml CCF4-AM dye solution in phenol-free DMEM/2% FBS was added to the cells. The CCF4-AM dye solution was prepared using a LiveBLazer FRET-B/G loading kit (Life Technologies) according to the alternative protocol recommended by the manufacturer. After incubation for 12–14 h at 11 °C in an ECHOterm chilling incubator (Torrey Pines Scientific), the cells were washed 3 × with PBS, detached with Versene (Life Technologies), fixed in 2% paraformaldehyde/PBS, and analysed on a Becton Dickinson LSR II flow cytometer. Samples were excited with a 405-nm violet laser, and fluorescence emission was measured in the Pacific Blue channel (450/50-nm filter) and in the AmCyan channel (525/20-nm filter).

49. Rodenburg, C. M. *et al.* Near full-length clones and reference sequences for subtype C isolates of HIV type 1 from three different continents. *AIDS Res. Hum. Retroviruses* **17**, 161–168 (2001).
50. Collman, R. *et al.* An infectious molecular clone of an unusual macrophage-tropic and highly cytopathic strain of human immunodeficiency virus type 1. *J. Virol.* **66**, 7517–7521 (1992).
51. Dorfman, T., Popova, E., Pizzato, M. & Gottlinger, H. G. Nef enhances human immunodeficiency virus type 1 infectivity in the absence of matrix. *J. Virol.* **76**, 6857–6862 (2002).
52. Dorfman, T., Mammano, F., Haseltine, W. A. & Gottlinger, H. G. Role of the matrix protein in the virion association of the human immunodeficiency virus type 1 envelope glycoprotein. *J. Virol.* **68**, 1689–1696 (1994).
53. Akagi, T., Shishido, T., Murata, K. & Hanafusa, H. v-Crk activates the phosphoinositide 3-kinase/AKT pathway in transformation. *Proc. Natl Acad. Sci. USA* **97**, 7290–7295 (2000).
54. Dettenhofer, M. & Yu, X. F. Highly purified human immunodeficiency virus type 1 reveals a virtual absence of Vif in virions. *J. Virol.* **73**, 1460–1467 (1999).
55. Chesebro, B., Wehrly, K., Nishio, J. & Perryman, S. Macrophage-tropic human immunodeficiency virus isolates from different patients exhibit unusual V3 envelope sequence homogeneity in comparison with T-cell-tropic isolates: definition of critical amino acids involved in cell tropism. *J. Virol.* **66**, 6547–6554 (1992).
56. Kelstrup, C. D., Young, C., Lavalley, R., Nielsen, M. L. & Olsen, J. V. Optimized fast and sensitive acquisition methods for shotgun proteomics on a quadrupole orbitrap mass spectrometer. *J. Proteome Res.* **11**, 3487–3497 (2012).
57. Accola, M. A., Strack, B. & Gottlinger, H. G. Efficient particle production by minimal gag constructs which retain the carboxy-terminal domain of human immunodeficiency virus type 1 capsid-p2 and a late assembly domain. *J. Virol.* **74**, 5395–5402 (2000).
58. Abacioglu, Y. H. *et al.* Epitope mapping and topology of baculovirus-expressed HIV-1 gp160 determined with a panel of murine monoclonal antibodies. *AIDS Res. Hum. Retroviruses* **10**, 371–381 (1994).
59. Schmittgen, T. D. & Livak, K. J. Analyzing real-time PCR data by the comparative C(T) method. *Nature Protocols* **3**, 1101–1108 (2008).



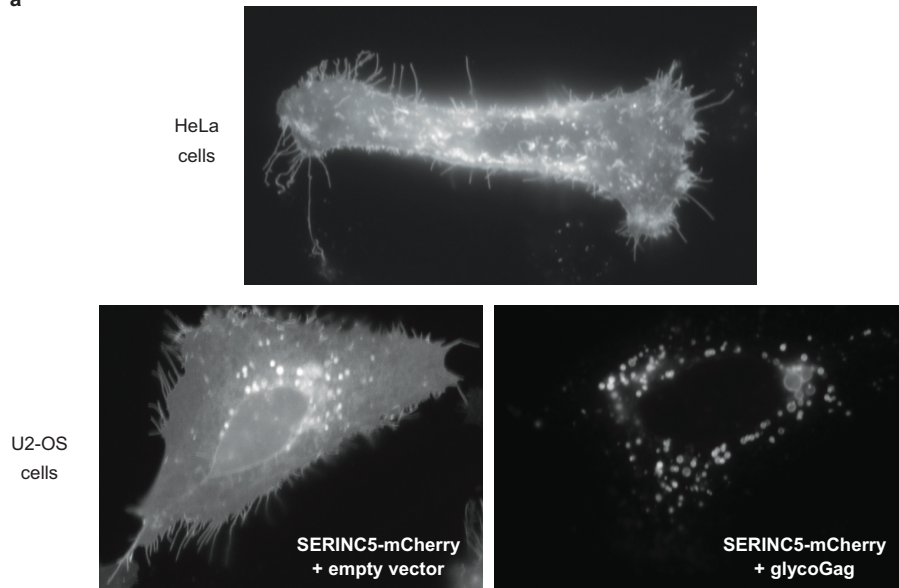
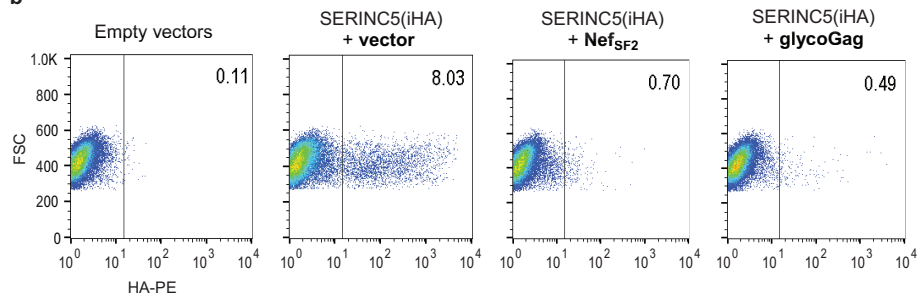
^aProteins identified with at least 5% coverage both in fraction 8 and in fraction 9 are shown

Extended Data Figure 1 | Identification of SERINC3 as a candidate target of Nef and glycoGag. **a**, Anti-HIV-1 CA immunoblot of Nef⁺, Nef⁻ and glycoMA⁺ HIV-1 virions collected from the indicated fractions of OptiPrep gradients. **b**, Proteins identified by mass spectrometry in Nef⁻ but not in Nef⁺ or glycoMA⁺ virion lysates. The data are from two independent experiments.



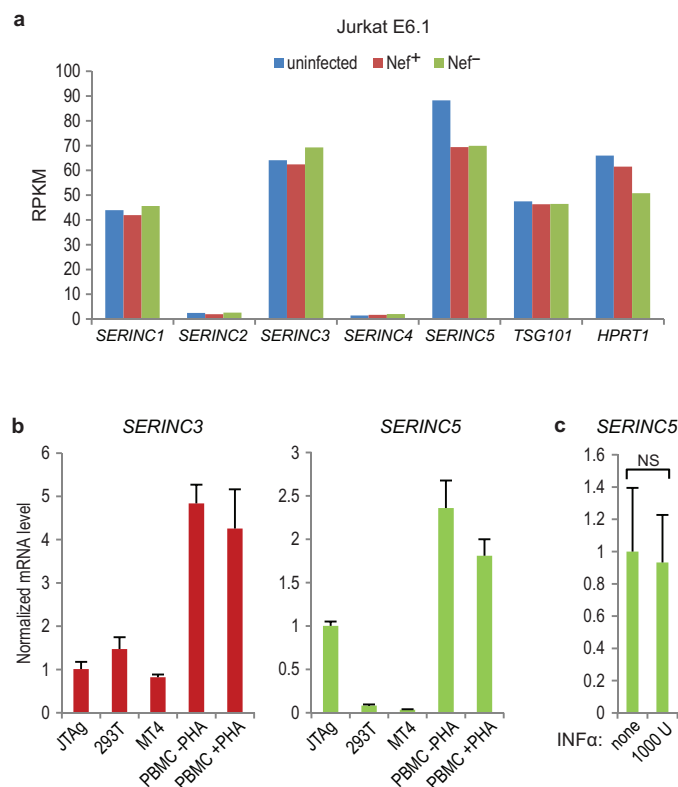
Extended Data Figure 2 | MLV glycoGag inhibits the incorporation of SERINC3 and SERINC5 into HIV-1 virions. **a, b**, Western blots showing the effects of wild-type or mutant glycoMA on the incorporation of SERINC3–HA (**a**) or SERINC5–HA (**b**) into Nef[−] HIV-1 virions. The NL4-3/glycoMA

proviral construct expresses untagged glycoMA *in cis*. In all other cases, HA-tagged (**a**) or Flag-tagged (**b**) glycoMA proteins were expressed *in trans*. The white bands marked by asterisks are caused by co-migrating HIV-1 Pr55^{gag}. Both experiments were performed twice.

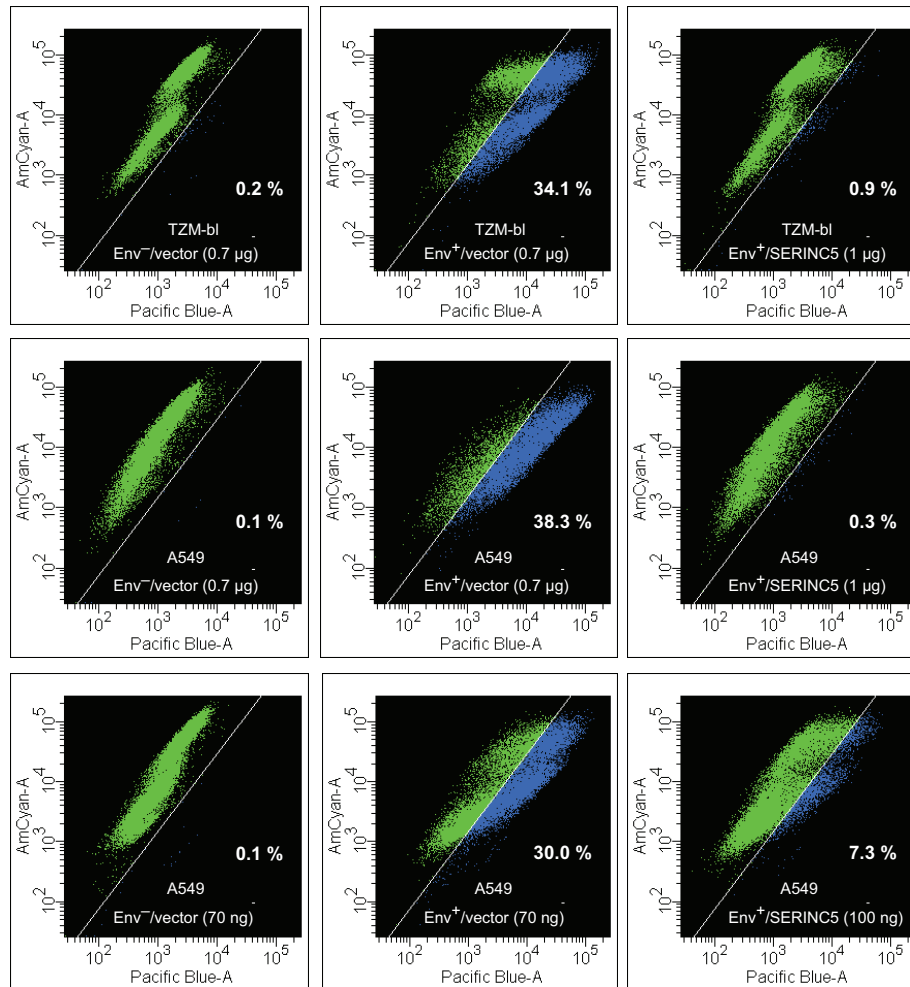
a**b**

Extended Data Figure 3 | Nef and glycoGag downregulate SERINC5 from the cell surface. **a**, SERINC5 re-localizes from the plasma membrane to perinuclear vesicles in the presence of glycoGag. HeLa or U2-OS cells transiently expressing SERINC5-mCherry alone or together with glycoGag were examined by live-cell fluorescence microscopy. **b**, Nef and glycoGag both

downregulate SERINC5. JTAG cells transiently expressing SERINC5(iHA), either alone or together with Nef_{SF2} or glycoGag, were surface-stained with anti-HA antibody and analysed by flow cytometry. Per cent fractions of cells expressing SERINC5(iHA) on the surface are indicated. This experiment was performed twice.

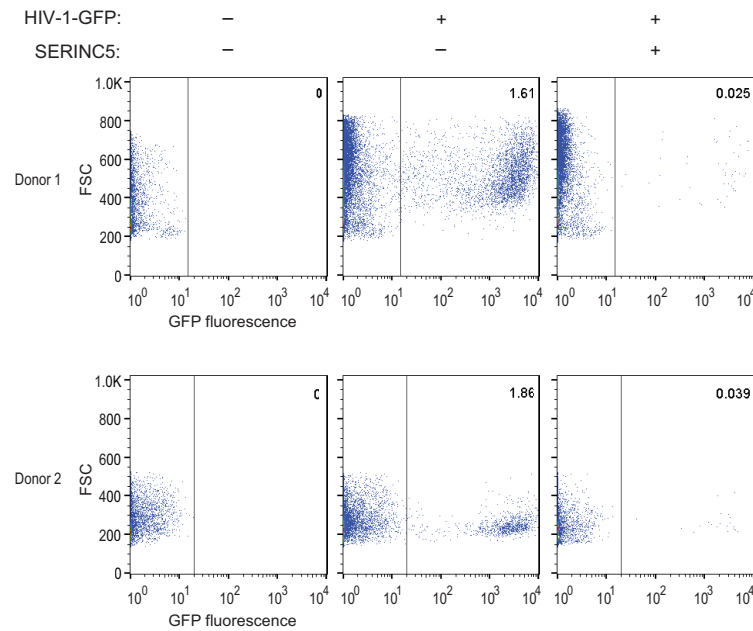


Extended Data Figure 4 | *SERINC* mRNA expression levels. **a**, Expression of *SERINC* family members in uninfected and HIV-infected Jurkat E6.1 cells. RNA was extracted at the peak of infection with wild-type (Nef⁺) or Nef⁻ HIV-1_{NL43}, and gene expression was quantified by RNA-seq as reads per kilobase of coding sequence per million reads (RPKM) ($n = 1$). The HIV-1 budding factor *TSG101* and the housekeeping gene *HPRT1* are included for comparison. **b**, Levels of *SERINC3* and *SERINC5* mRNA (arbitrary units) in cell lines and primary cells, as measured by qRT-PCR ($n = 3$). PBMC were left unstimulated or stimulated with $0.5 \mu\text{g ml}^{-1}$ phytohemagglutinin (PHA) and 20 U ml^{-1} IL-2 for 2 days. **c**, *SERINC5* mRNA expression is not induced by INF- α . PBMC were left untreated or treated with $1,000 \text{ U ml}^{-1}$ human INF- α 2a (PBL Assay Science) for 14 h ($n = 2$). Data are mean and s.d. NS, not significant ($P > 0.05$) two-tailed unpaired t -test.



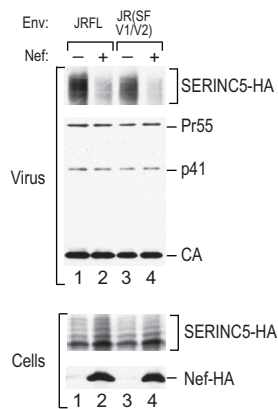
Extended Data Figure 5 | Exogenous SERINC5 inhibits the fusion of progeny virions with target cells. TZM-bl or A549/CD4/CXCR4 cells were exposed to equal amounts of virus containing BlaM-Vpr, and fusion was analysed by measuring the Env-dependent increase in blue fluorescence using multiparameter flow cytometry. Virions were produced in 293T cells

transfected with an Env⁻ HIV-1 provirus, a vector expressing Env_{HXB2} (Env⁺) or a frameshift mutant (Env⁻), a vector expressing BlaM-Vpr, and a vector expressing SERINC5 (1 µg or 100 ng) or an equimolar amount of the empty vector (0.7 µg or 70 ng). The percentage of cells displaying increased blue fluorescence is indicated.

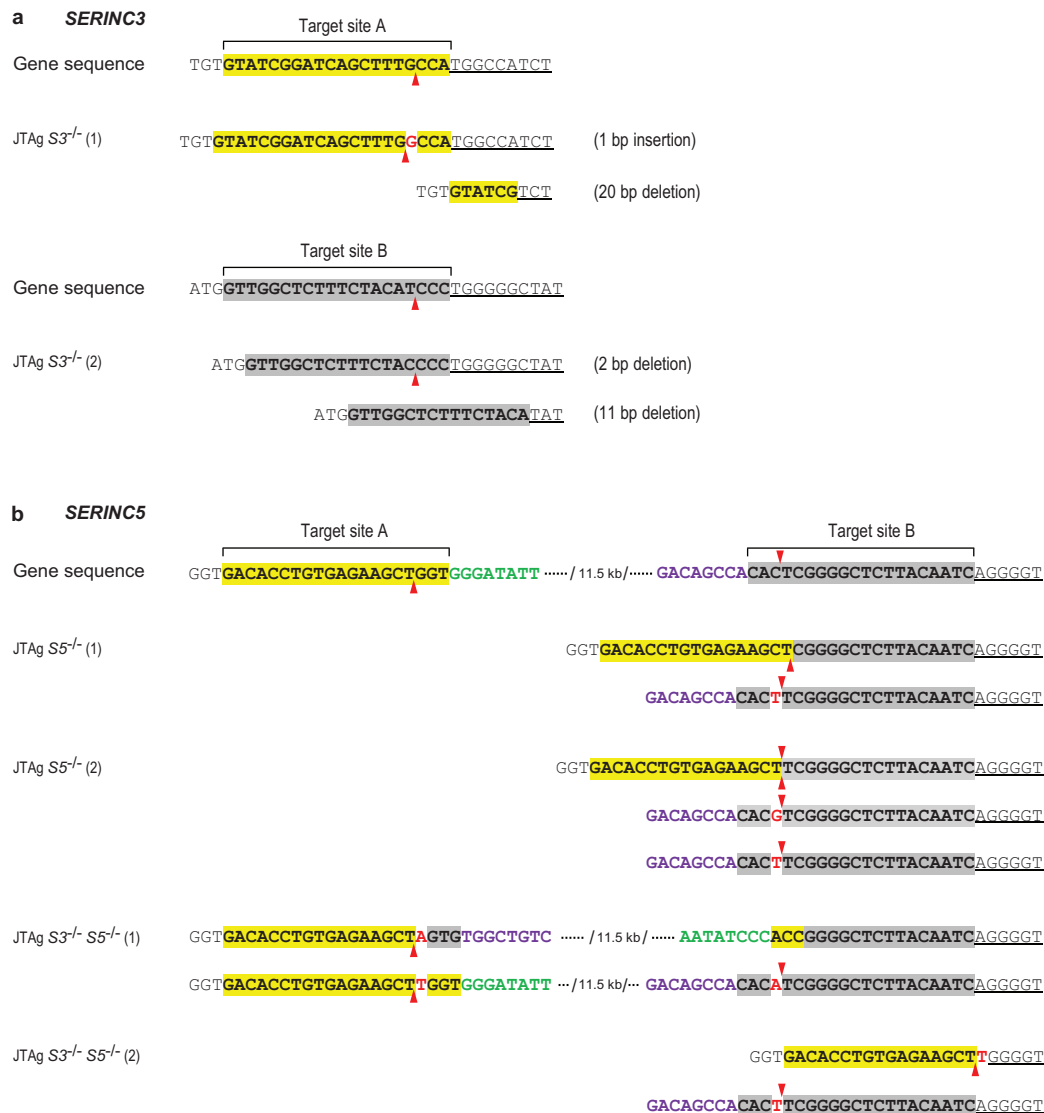


Extended Data Figure 6 | Exogenous SERINC5 reduces the infectivity of Nef^- HIV-1 progeny virions for primary target cells. In two independent experiments, PHA-stimulated PBMC from different donors were infected with

equal amounts of single-cycle GFP-HIV-1 virions produced in 293T cells in the absence or presence of exogenous SERINC5. Per cent fractions of infected (GFP-positive) cells are indicated.



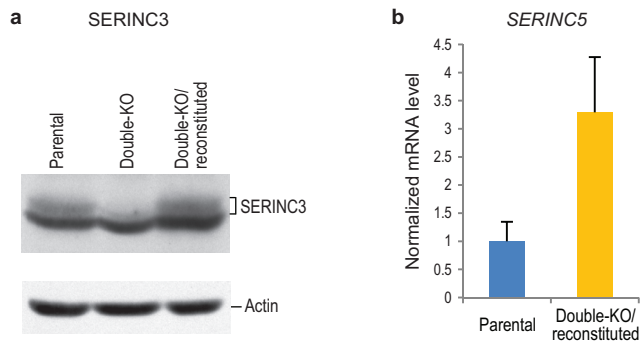
Extended Data Figure 7 | SERINC5 incorporation into HIV-1 virions that differ in Nef responsiveness. Recombinant virions were produced in 293T cells co-transfected with the HXB/Env⁻/Nef⁻ provirus and vectors expressing the poorly Nef-responsive Env_{JRFL} or the highly Nef-responsive JR(SF V1/V2) Env chimaera, along with a vector expressing SERINC5-HA. Empty pBJ5 vector or a version expressing HA-tagged Nef_{97ZA012} was also co-transfected. SERINC5-HA in purified virions was detected by western blotting. This experiment was performed twice.



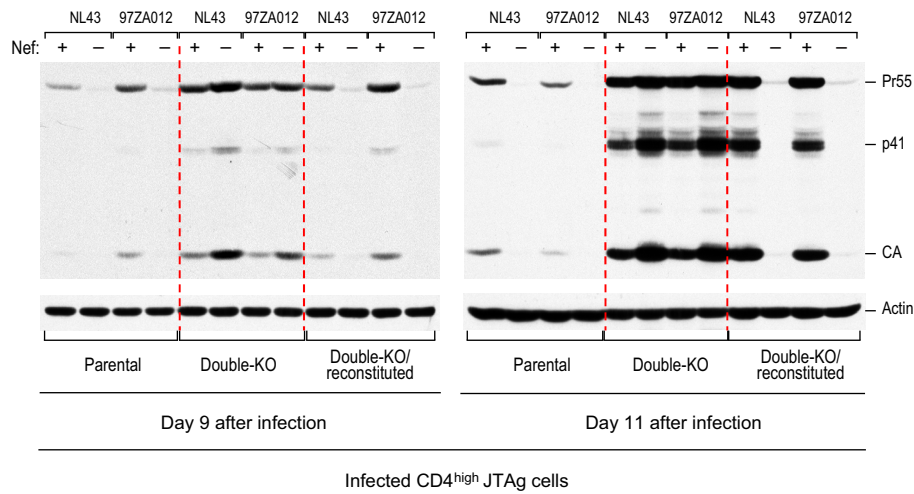
Extended Data Figure 8 | Characterization of JTAg knockout cells.

a, Mutant *SERINC3* alleles identified in *SERINC3* knockout clones. **b**, Mutant *SERINC5* alleles identified in *SERINC5* knockout and *SERINC3/5* double-knockout clones. The single-guide RNA (sgRNA) target sites are highlighted, and the predicted Cas9 target sites are indicated by arrowheads. Inserted

nucleotides are in red. One of the two mutated *SERINC5* alleles in JTAg *S3^{-/-} S5^{-/-}* (1) cells has an inversion between sgRNA target sites A and B. JTAg *S5^{-/-}* (2) cells contain three mutated *SERINC5* alleles. All mutations cause frameshifts and/or large deletions of coding sequence. No wild-type alleles were detected in any of the knockout clones.



Extended Data Figure 9 | SERINC3 and SERINC5 expression levels in reconstituted double-knockout cells. **a**, SERINC3 protein levels in parental, double-knockout, and reconstituted double-knockout JTag cells were compared by western blotting. SERINC3 migrated close to a prominent background band that was also recognized by the anti-SERINC3 antibody. **b**, SERINC5 mRNA levels in parental and reconstituted double-knockout JTag cells were compared by qRT-PCR ($n = 3$).



Extended Data Figure 10 | Effects of *SERINC* knockout and reconstitution on HIV-1 replication. Parental, double-knockout and SERINC3+SERINC5-reconstituted double-knockout CD4^{high} JTAG cells were analysed by

immunoblotting with anti-HIV CA at days 9 and 11 after infection with equal amounts (2 ng ml^{-1} p24) of HIV-1_{NL43} encoding either wild-type or disrupted versions of Nef_{NL43} or Nef_{97ZA012}.

Glycine receptor mechanism elucidated by electron cryo-microscopy

Juan Du^{1*}, Wei Lü^{1*}, Shenping Wu², Yifan Cheng² & Eric Gouaux^{1,3}

The strychnine-sensitive glycine receptor (GlyR) mediates inhibitory synaptic transmission in the spinal cord and brainstem and is linked to neurological disorders, including autism and hyperekplexia. Understanding of molecular mechanisms and pharmacology of glycine receptors has been hindered by a lack of high-resolution structures. Here we report electron cryo-microscopy structures of the zebrafish $\alpha 1$ GlyR with strychnine, glycine, or glycine and ivermectin (glycine/ivermectin). Strychnine arrests the receptor in an antagonist-bound closed ion channel state, glycine stabilizes the receptor in an agonist-bound open channel state, and the glycine/ivermectin complex adopts a potentially desensitized or partially open state. Relative to the glycine-bound state, strychnine expands the agonist-binding pocket via outward movement of the C loop, promotes rearrangement of the extracellular and transmembrane domain 'wrist' interface, and leads to rotation of the transmembrane domain towards the pore axis, occluding the ion conduction pathway. These structures illuminate the GlyR mechanism and define a rubric to interpret structures of Cys-loop receptors.

Neurotransmitter-gated ion channels mediate fast excitatory and inhibitory signal transduction in the central nervous system (CNS) by controlling ion flux through neuronal cell membranes in response to the binding of a wide range of neurotransmitters¹. Glycine, a major inhibitory transmitter in the CNS², exerts its inhibitory effect on the glycine receptor (GlyR), a postsynaptic ligand-gated channel receptor, opening a chloride-permeable pore that, in turn, leads to hyperpolarization of the membrane potential and inhibition of neuronal firing^{3–6}. GlyRs mediate neurotransmission throughout the spinal cord and brain stem and control a wide range of motor and sensory functions including vision and audition^{6–8}. Heritable mutations of human GlyR are the major cause of a rare neurological disorder, hyperekplexia (startle disease)^{9,10}.

Strychnine, the notoriously toxic and complex alkaloid¹¹, is a potent competitive GlyR antagonist that locks the receptor in a closed state, precluding chloride permeation. Used to facilitate receptor isolation¹², and exploited to disentangle glycine-induced synaptic currents, strychnine acts at the intersubunit, canonical neurotransmitter site¹³. Glycine binds at the same site as strychnine yet promotes channel opening, allowing permeation of chloride ions through an anion conductive pathway with an estimated diameter of 5.2–6.0 Å^{14,15}. Small molecules and ions, including the macrocyclic lactones ivermectin and related avermectins, modulate the gating activity of GlyRs by potentiating glycine-induced currents by an allosteric mechanism¹⁶. Despite the important roles of GlyRs in the CNS and their prominence in neuroscience, mechanisms to describe the action of strychnine, glycine and ivermectin in terms of atomic structure have proven elusive.

GlyRs belong to the superfamily of Cys-loop receptors that includes the cation-selective nicotinic acetylcholine receptor (nAChR) and the serotonin type-3 receptor (5-HT₃R), as well as the anion-selective GABA type A receptor (GABA_AR)^{17,18}. As a result of landmark studies¹, Cys-loop receptors have been studied because of their prominent roles in the nervous system and because they are targets of scores

of natural products and synthetic agents, from curare to valium. More recently, high-resolution structures of the prokaryotic pentameric ligand-gated ion channels, GLIC and ELIC^{19–22}, and a chimaeric GLIC–GlyR protein termed Lily²³, as well as the eukaryotic nAChR^{24,25}, GluCl^{26,27}, 5-HT₃R²⁸, and GABA_AR²⁹ have been elucidated. Molecular understanding of eukaryotic Cys-loop receptors, however, is largely based on comparisons between different receptors^{30,31} due to the challenge of capturing a single receptor in multiple functional states. To reveal the molecular interplay between competitive antagonists, agonists or allosteric modulators and ion channel gating, we determined GlyR structures in complex with strychnine, glycine, or glycine and ivermectin (referred to here as glycine/ivermectin) using single-particle electron cryo-microscopy (cryo-EM).

Structure determination and refinement

The three-dimensional reconstructions of the strychnine-, glycine-, and glycine/ivermectin-bound structures have estimated resolutions of 3.9, 3.9, and 3.8 Å, respectively (Extended Data Figs 1–3), and are of sufficient quality to allow modelling of almost the entire receptor (Fig. 1 and Extended Data Fig. 4). The density for strychnine can be recognized in the strychnine-bound form, and is located at the intersubunit neurotransmitter-binding pocket. By contrast, density for glycine is not discernable in either the glycine- or the glycine/ivermectin-bound forms. Characterized by the distinctive triangular shape of its macrocyclic lactone, the density of ivermectin is unambiguous, and is found wedged between transmembrane domain (TMD) subunit interfaces. The pore-lining M2 helix is best resolved in the glycine/ivermectin- and strychnine-bound structures (Extended Data Fig. 4c), although three (Ala326–Thr328) and two (Gly327–Thr328) residues are not visible in the M3–M4 loop of the strychnine- and glycine-bound reconstructions, respectively (Extended Data Fig. 4d). The final structures have excellent stereochemistry (Extended Data Table 1) and correlate well with the respective density maps (Extended Data Figs 1–3).

¹Vollum Institute, Oregon Health & Science University, 3181 SW Sam Jackson Park Road, Portland, Oregon 97239, USA. ²Department of Biochemistry and Biophysics, University of California San Francisco, 600 16th Street, San Francisco, California 94158, USA. ³Howard Hughes Medical Institute, Oregon Health & Science University, 3181 SW Sam Jackson Park Road, Portland, Oregon 97239, USA.

*These authors contributed equally to this work.

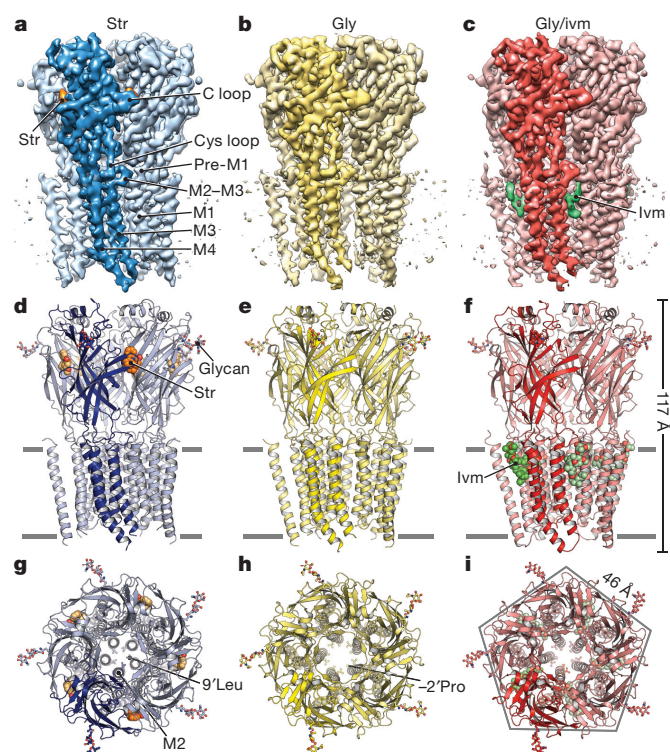


Figure 1 | Glycine receptor architecture. **a–c**, The three-dimensional reconstruction maps, viewed parallel to the membrane (strychnine (str)-bound in blue, glycine (gly)-bound in yellow, and glycine/ivermectin (ivm)-bound in red). One subunit is highlighted. The densities for strychnine and ivermectin are orange and green, respectively. **d–f**, Cartoon representations of the corresponding models of reconstructions shown in **a–c**, viewed parallel to the membrane plane. The Asn-linked carbohydrate and associated Asn54 residue are shown in stick representation. **g–i**, Views of the structures from the extracellular side of the membrane. Residues $-2'$ Pro (Pro266) and $9'$ Leu (Leu277) reside on the pore-lining M2 helix.

Overall architecture

The three GlyR structures have urn-like architectures, similar to the AChR²⁵ and other Cys-loop receptors, with the pentameric assemblage of subunits surrounding a central fivefold symmetric pore axis. Each subunit has the shape of an upright left forearm, wearing a mitten (Extended Data Fig. 5). The β -strands of the extracellular domain (ECD) form the palm and the hallmark C loop emerges as the thumb, poised on the back of the hand of an adjacent subunit. A 'wrist-like' cuff connects the ECD and TMD and is buttressed by 'ligaments' that include the pre-M1 linker and loops of the ECD and TMD. We envision the TMD as the 'forearm', consisting of helices M1–M4. With the open palm and thumb of one mitten packing against the back of the hand of the neighbour, together with multiple interactions via the TMD forearms, these interactions knit the pentamer together (Fig. 1d–f).

The outline of the strychnine–GlyR TMD is approximately perpendicular to the membrane (Fig. 1a, d). By contrast, in the glycine–GlyR complex, the extracellular half of TMD is wider than the intracellular half (Fig. 1b, e), reminiscent of a truncated cone. In the glycine/ivermectin–GlyR structure, the intracellular halves of the TMs move closer to each other (Fig. 1c, f). Thus, the pore in the strychnine-bound form is constricted throughout and the pore-lining M2 helices are perpendicular to the membrane, in comparison with the other two forms (Fig. 1g–i). In the glycine-bound form, the extracellular opening is wider than in the strychnine complex because the M2 helices tilt and undergo an anticlockwise rotation around the pore axis (viewed from the extracellular side). The binding of ivermectin promotes a clockwise rotation of the intracellular half of the M2,

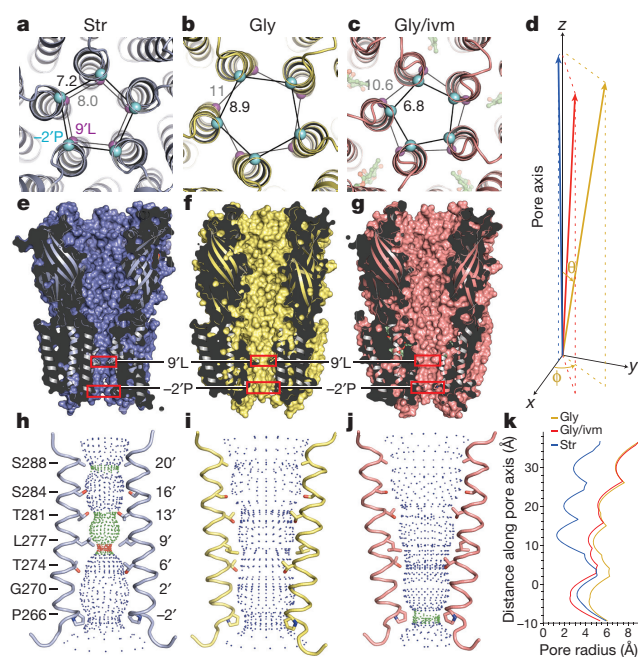


Figure 2 | The ion channel. **a–c**, Two sites of pore constriction are at $-2'$ Pro and $9'$ Leu, viewed from the cytoplasmic side, with their C_{α} in cyan and magenta spheres, respectively. Distances are in ångströms. **d**, Plot of the pore-lining M2 as represented by the vector connecting $-2'$ Pro C_{α} and $9'$ Leu C_{α} , with $-2'$ Pro C_{α} as the origin, the tilt angle θ and the rotation angle ϕ relative to the pore axis. The ϕ of strychnine-bound GlyR is set to zero. **e–g**, Sagittal 'slice' views along the pore axis. **h–j**, Shape and size of the ion permeation pathway. M2 of two subunits are shown as ribbon representation and the side chains of the pore-lining residues are shown in sticks. Blue, green, and red spheres define radii of >3.3 Å, 1.8 – 3.3 Å, and <1.8 Å, respectively. **k**, Plot of pore radii as a function of distance along the pore axis. The C_{α} position of $0'$ Arg is set to zero.

relative to the glycine-bound form, constricting the intracellular opening of the pore but leaving the extracellular entry as expanded as in the glycine-bound structure.

Ion channel pore

Inspection of the three GlyR structures demonstrates that there are two physical sites of constriction within the ion channel pore located at $-2'$ Pro (Pro266) and $9'$ Leu (Leu277), highly conserved residues that reside on the M2 helix and are located near the cytoplasmic side and the middle of the helix, respectively. The descending order of pore cross-sectional area, as estimated by the distance between adjacent C_{α} atoms of $-2'$ Pro residues, is glycine, strychnine, glycine/ivermectin, while at the position of the $9'$ Leu, the order is glycine, glycine/ivermectin, strychnine (Fig. 2a–c). The ion channel in the strychnine-bound form is the narrowest, with a radius of ~ 1.4 Å at $9'$ Leu, too narrow for passage of a dehydrated chloride ion with an ionic radius of 1.8 Å³². In the glycine- and glycine/ivermectin-bound structures, $9'$ Leu rotates away from the five-fold axis, expanding the ion channel pore to a radius of 4 – 5 Å. In the glycine-bound form, the smallest radius is 4.4 Å at $-2'$ Pro, allowing permeation of hydrated chloride ions, with an ionic radius of 3.3 Å. Thus the strychnine-bound form is an antagonist-bound closed non-conducting state and the glycine-bound form is an agonist-bound open conducting state (Fig. 2e–j).

To define the structural mechanism underlying the changes in dimension and shape of the ion channel pore, we measured the 'tilt' (θ) and rotation angles (ϕ) of the pore-lining M2 helices relative to the pore axis (Fig. 2d). The tilt angle θ is approximately the same in the glycine and glycine/ivermectin complexes and larger than in the strychnine form, thus opening the extracellular 'halves' of the ion channel pores in the glycine- and glycine/ivermectin-bound forms nearly twofold

the width of the strychnine-bound state (Fig. 2h–k). Furthermore, by measuring the rotation angle ϕ of the M2 helix, we see that the M2 helices have undergone anticlockwise rotations of 49° and 22° in the glycine- and glycine/ivermectin-bound states, respectively, relative to the strychnine-bound form, when viewed from the extracellular side of the membrane. Considering the tilt and rotation angles together, the M2 helix undergoes an outward rotation from the strychnine- to glycine-bound forms, enlarging both constriction sites. By contrast, the relatively smaller rotation in the glycine/ivermectin-bound state leads to a distinct repositioning of the cytoplasmic end of the M2 helix, as defined by measurements at $-2'$ Pro, resulting in a constriction of the pore to a radius of 2.5 Å, which is smaller than a hydrated chloride ion (Fig. 2k).

To ground our interpretations of the GlyR structures in the context of physiological function, we performed two-electrode voltage clamp electrophysiology (TEVC) experiments. At intermediate (Fig. 3g) and saturating (Fig. 3f) glycine concentrations, current decays are small and slow, consistent with the notion that the GlyR_{EM} construct is not prone to desensitization and that the glycine-bound structure is an open conducting state. Strychnine antagonizes the glycine currents (Fig. 3g), in accord with the conclusion that the strychnine-bound structure is a closed non-conducting state. In the presence of glycine and ivermectin, we observe initial potentiation followed by pronounced decay (Fig. 3h). In addition, block of the glycine/ivermectin current by picrotoxinin, an open channel blocker, is only $\sim 20\%$ on peak or steady-state currents. Thus, even though ivermectin initially enhances glycine-induced currents, we propose that the glycine/ivermectin-bound form represents an agonist/allosteric modulator-bound desensitized state or a partially open state with reduced susceptibility to picrotoxinin block. This is consistent with recent studies which suggest that desensitization in ELIC³³ involves constriction at the intracellular entrance of the pore and that the picrotoxinin-binding site in GlyR spatially overlaps the desensitization gate³⁴.

We can define a rubric for extant Cys-loop receptor structures by dividing the receptors into three groups (Extended Data Fig. 7a–d): (1) closed: strychnine–GlyR, apo–GluCl, 5-HT₃R, nAChR; (2) partially open or desensitized-like: glycine/ivermectin–GlyR, glutamate/ivermectin–GluCl, GABA_AR; (3) open: glycine–GlyR. In group 1, the pores are restricted by 9°Leu. However, they possess distinct M2 conformations as indicated by the rotation angle (ϕ), perhaps reflective of their distinct states: antagonist/closed and resting/closed. In group 2, glycine/ivermectin–GlyR and glutamate/ivermectin–GluCl share similar pore profiles and M2 conformations, suggesting they

represent similar physiological states, perhaps desensitized-like or partially open low-conductance states³⁵. Despite a tighter restriction at the intracellular entrance, the pore properties of the GABA_AR resemble that of glycine/ivermectin–GlyR. This is consistent with the suggestion that the GABA_AR crystal structure reflects an agonist-bound desensitized state²⁹. In contrast, the unique large pore size and distinct M2 orientation distinguish the glycine–GlyR structure and are consistent with it defining a fully open ion-conducting state.

Neurotransmitter-binding site

The neurotransmitter-binding site of Cys-loop receptors is located at the interface of two adjacent subunits, surrounded by the C loop ‘thumb’ from the (+) side and the ‘back of the mitten’ β -strands on the (–) side. Strychnine binds with a dissociation constant (K_d) of 98 nM (Fig. 3c) and in the strychnine–GlyR structure, we observe strong density in the neurotransmitter-binding site (Fig. 3a). The almond-shaped density hews to the shape of a strychnine molecule, whose ring I forms the tip and rings III to VII form the bulbous bottom. In this pose, the relative positions of strychnine and the side chain of the conserved Tyr218 are in harmony with the crystal structure of the AChBP–strychnine complex³⁶. A possible hydrogen bond between the carbonyl oxygen of strychnine and Arg81, proximity of the same carbonyl to Thr220 (C loop), and sandwiching of strychnine’s aromatic rings by Phe79 and Tyr218 further enhance interactions between the antagonist and receptor (Fig. 3a, b). In addition, the mutation Tyr177Ala results in strychnine insensitive GlyRs, whereas Tyr177Phe does not³⁷. Despite no direct interaction to strychnine, Tyr177 sterically restricts the B loop close to the neurotransmitter-binding pocket, thus explaining how a change in residue volume could indirectly perturb the binding site.

Glycine robustly activates the receptor with minimal or partial current decay upon prolonged application of agonist at intermediate or saturating concentrations, respectively (Fig. 3f, g) and yields an EC₅₀ of 0.26 mM. Density associated with glycine is not discernible in either the glycine- or glycine/ivermectin-bound structures due to its low molecular mass. Nevertheless, mutational analysis indicates that strychnine binding determinants such as Thr220 and Tyr218 partially overlap with those of glycine^{38–40}. In particular, the Tyr218Phe mutant is reported to produce a 480-fold reduction in glycine affinity⁴¹. By comparing the ECDs of the glycine- and strychnine-bound structures, we observe that the overall shape of the β -sheet forming the (–)-half of the binding pocket remain essentially unchanged (Fig. 4a). In contrast, Arg81, which is involved in glutamate binding in GluCl²⁶, exhibits

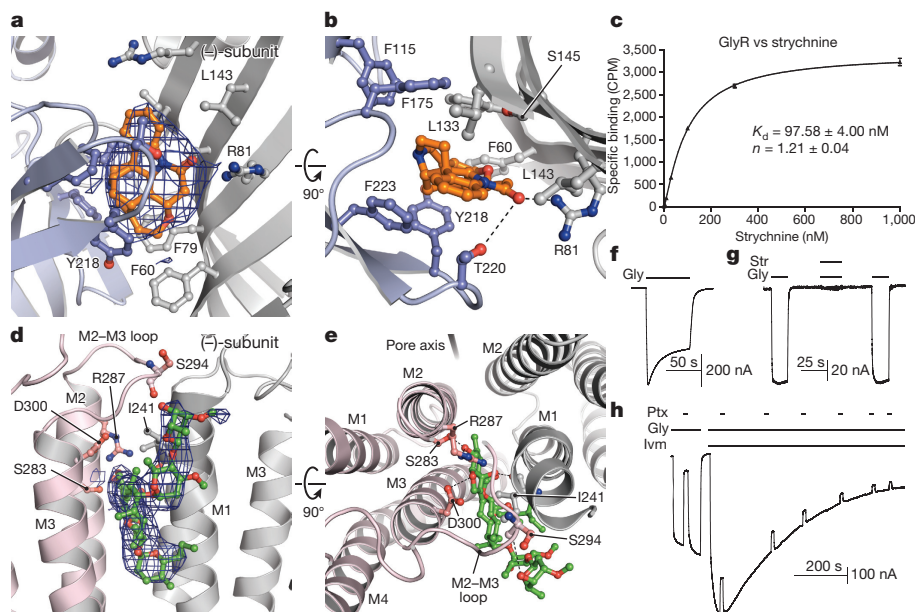


Figure 3 | Strychnine and ivermectin bind at subunit interfaces. **a**, **b**, Strychnine binding site (+, light blue) (–, grey); views are parallel to the membrane (**a**) or from the extracellular side (**b**). Density for strychnine (blue mesh) is contoured at 7σ . **c**, Saturation binding of ³H-strychnine to the GlyR_{EM} construct. Results are the mean of three biological replicates and the error bars represent s.e.m. **d**, **e**, Ivermectin binds at the TMD intersubunit interface. Views are parallel to the membrane (**d**) or the extracellular side (**e**). **f**, Activation of GlyR currents by 10 mM glycine determined by TEVC. **g**, Strychnine (1 μ M) inhibits glycine-induced (0.3 mM) currents. **h**, Effect of picrotoxinin (ptx, 1 mM) on (0.3 mM) glycine-induced current. Picrotoxinin inhibits $\sim 80\%$ of glycine-induced current. Ivermectin (5 μ M) potentiates the GlyR current and causes a slow desensitization. The recordings shown are representative of three independent experiments.

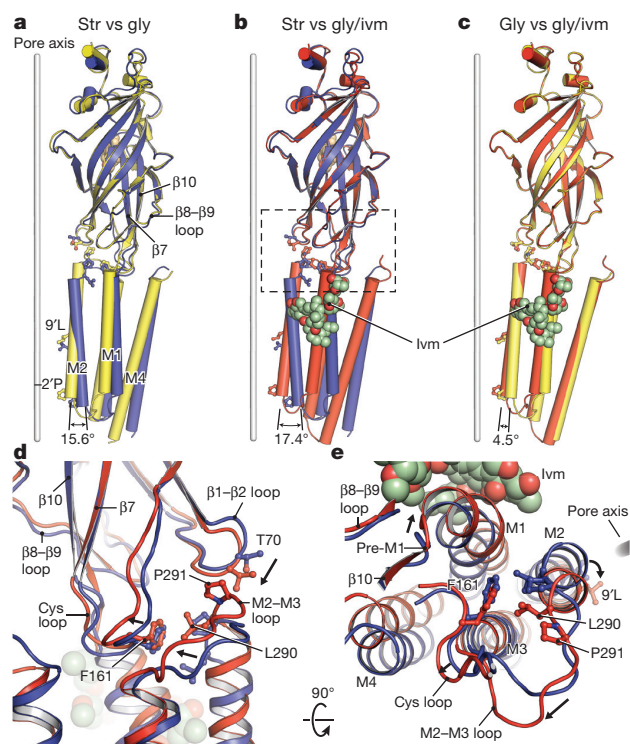


Figure 4 | Conformational changes within an individual subunit.

Strychnine-, glycine-, and ivermectin/glycine-bound states are in blue, yellow, and red, respectively. **a–c**, Superimposition of the three GlyR structures using the ECD (residues 1–235), showing the motion of the TMD. Relative rotation angles of the pore-lining M2 are indicated. **d, e**, Conformational changes in the ECD–TMD interface upon transition from the glycine- (or glycine/ivermectin-) to the strychnine-bound states are shown viewed parallel to the membrane (**d**) and from the extracellular side (**e**). The displacement of the β8–β9 loop leads to a rotation of pre-M1/M1, pushing the lower half of M2 towards the pore axis; meanwhile, this displacement repositions the Cys loop through β10, which results in the coupling of the M2–M3 loop with the β1–β2 loop through the interaction between Pro291 and Thr70. Consequently, the upper half of M2 rotates outward.

substantial conformational changes (Fig. 5b, c). We propose that Arg81, Thr220 and Tyr218, together with Ser145, Phe175 and Phe223, directly participate in the binding of glycine^{41–43}.

A central concept of the relationship between ligand binding and channel gating in Cys-loop receptors is that the conformation of the C loop thumb is correlated to the functional state of the receptor, that is, the C loop is ‘open’ in the antagonist-bound closed channel state and ‘closed’ in the agonist-bound open channel state^{44–48}. To examine this hypothesis, we superimposed the (–)-subunit of our GlyR models with the glutamate/ivermectin–GluCl, apo–GluCl, GABA_AR, 5-HT₃R, and strychnine–AChBP structures and compared the size of their binding pockets by measuring the distance between the C loop and the right half of the pocket (Extended Data Fig. 7e). We observed that antagonist-bound strychnine–GlyR and strychnine–AChBP structures possess more open C loops and remarkably larger binding pockets than the agonist-bound structures. In addition, the glycine–GlyR structure shares a more similar pocket size and shape to the glutamate/ivermectin–GluCl than the apo–GluCl structure, consistent with glycine being bound in the glycine–GlyR structure.

Allosteric-binding site

Ivermectin is a GlyR allosteric modulator that potentiates glycine-induced currents (Fig. 3h). In the glycine/ivermectin-bound structure, the wedge-shaped ivermectin inserts into the TMD interface of two adjacent subunits and interacts with hydrophobic residues in M3 of the (+)-subunit and M1 of the (–)-subunit (Fig. 3d, e). It also probably

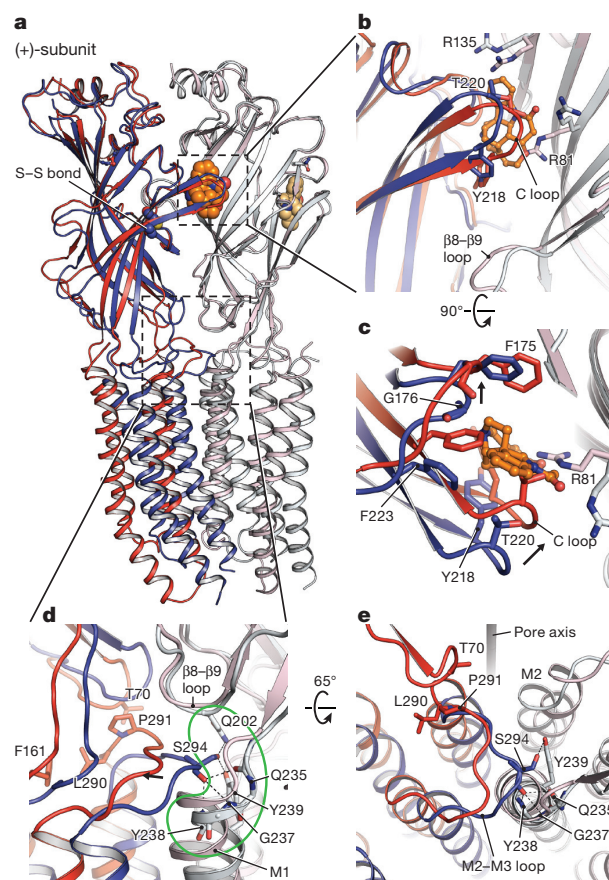


Figure 5 | Conformational differences at the subunit-subunit interface between agonist- and antagonist-bound states.

Strychnine- and glycine/ivermectin-bound states are in blue and red, respectively. The (–)-subunits are in corresponding light colours. **a**, Superimposition of the ECD of the (–)-subunits showing the relative movement of the (+)-subunits. **b, c**, Conformational changes of the neurotransmitter binding pocket, viewed parallel to the membrane (**b**) and from the extracellular side (**c**). The neurotransmitter-binding site expands in the strychnine-bound structure caused by repositioning of Arg135 and Arg81 in the (–)-subunit and by the opening of the C loop in the (+)-subunit. **d, e**, The coupling of structural rearrangements of the ECD–TMD interface between two adjacent subunits. In the strychnine-bound form, Ser294 in the M2–M3 loop of the (+)-subunit is inserted in the M1 N-cap in the (–)-subunit. Key residues interacting with Ser294 are highlighted in the green outline. Upon binding of glycine, the M2–M3 loop moves away from the N-cap. For clarity, the side chains of Gln235 and Tyr238 are not shown.

forms hydrogen bonds with several residues, one of which is Arg287, a residue that enables a direct interaction between ivermectin and the pore-lining M2 helix. Mutations of Arg287 to Gln or Leu in the human α1 GlyR subunit reduce the sensitivity of the receptor to ivermectin by a factor of 20 and these substitutions are the most frequent mutants associated with hyperekplexia (Extended Data Fig. 9c)^{5,16}. The potential interaction between Arg287 and ivermectin, together with the location of Arg287 on the M2 helix, explains why mutations at 287 perturb both ivermectin action and ion channel gating.

Unlike GluCl²⁶, ivermectin is not required for GlyR activation. To probe the structural basis for this difference in function, we compared the binding pockets in both receptors (Extended Data Fig. 6). Whereas ivermectin binds similarly to both receptors, there are several notable differences. First, Arg287 in GlyR interacts with ivermectin, yet the corresponding residue in GluCl²⁶ is an asparagine (Asn264) and the side chain is positioned further from ivermectin. Second, Val296 in the M2–M3 loop of GlyR is an isoleucine (Ile273) in GluCl, whose larger side chain prevents the O6-oxygen of ivermectin from approaching the M2–M3 loop. Third, Gly237 in the M1 and

Ala304 in the M3 of GlyR are Ser217 and Gly281 in GluCl, respectively. In GlyR, the M3 residue (Ala) is larger than the M1 residue (Gly), effectively ‘pushing’ ivermectin towards M1, creating a larger interaction surface area between ivermectin and M1 compared to M3. This situation is reversed in GluCl, however, as the M1 residue (Ser) is larger than the M3 residue (Gly). This is consistent with the experiments where the Ala304Gly substitution in the M3 strikingly enhanced the ivermectin sensitivity of human α_1 GlyR⁴⁹, probably by enlarging the ivermectin–M3 interface.

Signal transduction

By superimposing a single subunit from the strychnine- and glycine-bound states, we see that the β -sheets and flanking loops of the ECD behave like a malleable ‘palm’, undergoing a ‘flexing-like’ conformational change that is distributed throughout the entire domain, thus defying the demarcation of a simple local hinge (Fig. 4a). The superposition also reveals that binding of ivermectin does not cause significant conformational changes of the ECD and ECD–TMD interface (Fig. 4b, c). In the following comparison of the ECD and ECD–TMD interfaces, the glycine/ivermectin–GlyR structure, whose reconstruction is of better quality, is used (Figs 4d, e and 5). By contrast, the TMD behaves as an approximate rigid body upon comparison of the three structures (Extended Data Fig. 8). The M3–M4 loop is on the cytoplasmic side of the TMD, is truncated in this GlyR_{EM} construct, and may have an effect on the conformational changes of the TMD. Further studies are necessary to clarify the role of the M3–M4 loop in receptor gating. Nevertheless, when we analyse the ECD interface in the strychnine- and glycine-bound states by superposition using the ECD of the (–)-subunit (Fig. 5a), the ‘flexing’ of the β -sheets leads to prominent conformational changes in two critical regions—the interfaces between the ECDs that span the crevices between the orthosteric and allosteric agonist-binding sites, and the interfaces between the ECD and the TMD.

At the neurotransmitter-binding site, we observed an expansion in the strychnine-bound structure caused by displacement of Arg81 and Arg135 on the core of the β -sheet in the (–)-subunit and by the opening of the C loop, as indicated by the outward movement of Thr220 C α by 4 Å (Fig. 5b, c). The motion of the C loop is mainly a consequence of the movement of the entire ECD, as also observed in GluCl²⁷, together with a subtle ‘flexing’ at the loop tip. At the ECD–TMD interface, in the strychnine-bound closed state, there are extensive interactions between subunits at the interface near the ECD–TMD boundary⁵⁰, defined by interactions between the β 1– β 2 loop, Cys loop, and M2–M3 loop of the (+)-subunit with the β 1– β 2 loop, β 8– β 9 loop, and pre-M1/M1 of the (–)-subunit. There are also multiple contacts between subunits within the TMD, mediated by intersubunit interactions between the M1 and M3 helices, and between M2 helices. By contrast, in the glycine-bound state, the interactions between subunits at the ECD–TMD boundary have largely ruptured, simply due to an increase in the separation of subunits, and the contacts between the TMD have also diminished by more than 50% (540 Å² in strychnine–GlyR and 240 Å² in glycine–GlyR), also because the subunits have moved apart. Remarkably, the increase in the separation of the intersubunit M1–M3 interaction in the glycine-bound state provides the initial ‘indentation’ of the cavity that is occupied by ivermectin in the glycine/ivermectin complex.

To further understand how the conformational changes within the ECD are transduced to the TMD, we divided the M2 helix into upper and lower halves and followed their movements. On the one hand, the upper half of M2 couples to both the β 1– β 2 loop and Cys loop through the M2–M3 loop (Fig. 4d, e) in the glycine-bound form. The Cys loop further interacts with pre-M1/ β 10, the latter of which moves along with the β 8– β 9 loop. On the other hand, the lower half of M2 is connected to M1, which is coupled to the β 8– β 9 loop through the pre-M1 loop. Thus, the movement of M2 can be traced to the β 8– β 9 loop, a loop that, in turn, not only connects to the C loop covering

the binding pocket but also sits underneath the pocket and thus indirectly participates in the binding pocket of the adjacent subunit (Fig. 5b). In going from the strychnine- to glycine-bound states (Figs 4d, e and 5b, c), the C loop thumb switches from ‘open’ to ‘closed’ and the β 8– β 9 loop is displaced. This leads to a concomitant rotation of pre-M1 and M1, thus pushing the lower half of M2 towards the pore axis. Meanwhile, the β 8– β 9 loop repositions the Cys loop through β 10, which results in the interaction of the M2–M3 loop with the β 1– β 2 loop causing an outward rotation of the upper half of M2.

Particularly important are the interactions at the ECD–TMD interface, where the β 1– β 2 and Cys loop of the (+)-subunit, together with the β 8– β 9 loop and pre-M1/M1 of the (–)-subunit, interact with the crucial M2–M3 loop of the (+)-subunit (Fig. 5d, e). Here, in the strychnine-bound form, we suggest that Ser294 in the M2–M3 loop of the (+)-subunit stabilizes the channel in a closed conformation by capping the amino terminus of the M1 helix of the (–)-subunit (N-cap), making van der Waals contacts with residues in the pre-M1/M1 region and participating in interactions with residues at the β 8– β 9 loop tip. Upon binding of glycine, the M2–M3 loop moves away from capping the amino terminus of the M1 helix and instead interacts with the β 1– β 2 loop via Pro291 and Thr70, stabilizing the channel in an open form. Several startle disease mutants in the M1 helix and the M2–M3 loop cause spontaneous activation of the receptor, perhaps by disruption of the interactions between Ser294 and the M1 N-cap (Extended Data Fig. 9a, b)¹⁰.

Comparisons of the glycine- and glycine/ivermectin-bound structures reveal that binding of ivermectin causes the most notable changes at the cytoplasmic half of the TMD by ‘tilting’ the lower half of the M2 helix towards the pore lumen by 4.5° (Fig. 4c), contracting the intercellular opening of the ion channel pore at –2’Pro (Fig. 2j, k). However, the extracellular regions of the pore in the glycine- and glycine/ivermectin-bound states are similar, possibly constrained by interactions of ivermectin with the M2 helix and the M2–M3 loops.

Gating mechanism

The strychnine- and glycine-bound structures of the GlyR, elucidated free from constraints of a crystal lattice or Fab fragments, unambiguously define antagonist-locked/closed and agonist-activated/open states of the receptor, respectively, while the glycine and ivermectin complex is suggestive of an allosteric modulator-bound low conductance or possibly desensitized state. A mechanism of ligand-dependent gating emerges from analysis of these structures in which the palm of the ECD and the forearm of the TMD are coupled by a flexible joint at the ECD–TMD interface, a joint reinforced by a ligament-like cuff

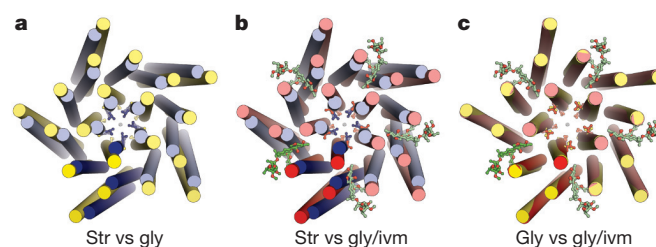


Figure 6 | Overall conformational changes of the TMD. a–c, Strychnine-, glycine-, and glycine/ivermectin-bound states are in blue, yellow, and red, respectively. Comparison between strychnine- and glycine–GlyR (a), strychnine- and glycine/ivermectin–GlyR (b), glycine-, and glycine/ivermectin–GlyR (c), viewed from the extracellular side. Side chains of –2’Pro and 9’Leu are shown in sticks to denote the change of pore sizes. In going from the strychnine- to the glycine-bound form, the TMD of each subunit undergoes an anticlockwise outward rotation, enlarging the pore size by pulling the side chains of 9’Leu and –2’Pro away from the channel axis. Binding of ivermectin to the glycine–GlyR causes a clockwise inward rotation of the TMD. As a result, although the extracellular half of the pore undergoes little change, the intracellular entrance shrinks.

composed of interacting ECD and TMD loops. Agonist binding closes the C loop thumb, promotes a concerted anticlockwise rotation around the pore axis (viewed from the extracellular side) of all 5 palm domains about an axis formed by finger-palm joints, and thus an iris-like expansion and anticlockwise rotation of the entire TMD 'forearms' (Fig. 6a–c and Supplementary Video 1). These structures not only allow us to describe a molecular mechanism for ligand-dependent gating in GlyRs, but also define a framework for interpreting the wealth of structural information on Cys-loop receptors and bacterial orthologues, which in turn will ground the search for new therapeutic agents on a solid structural foundation.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 17 March; accepted 23 June 2015.

Published online 7 September 2015.

- Katz, B. & Thesleff, S. A study of the 'desensitization' produced by acetylcholine at the motor end-plate. *J. Physiol. (Lond.)* **138**, 63–80 (1957).
- Aprison, M. H. & Werman, R. The distribution of glycine in cat spinal cord and roots. *Life Sci.* **4**, 2075–2083 (1965).
- Curtis, D. R., Hosli, L. & Johnston, G. A. Inhibition of spinal neurons by glycine. *Nature* **215**, 1502–1503 (1967).
- Werman, R., Davidoff, R. A. & Aprison, M. H. Inhibition of motoneurons by iontophoresis of glycine. *Nature* **214**, 681–683 (1967).
- Curtis, D. R., Hosli, L., Johnston, G. A. & Johnston, I. H. The hyperpolarization of spinal motoneurons by glycine and related amino acids. *Exp. Brain Res.* **5**, 235–258 (1968).
- Lynch, J. W. Molecular structure and function of the glycine receptor chloride channel. *Physiol. Rev.* **84**, 1051–1095 (2004).
- Galzi, J. L., Edelstein, S. J. & Changeux, J. The multiple phenotypes of allosteric receptor mutants. *Proc. Natl Acad. Sci. USA* **93**, 1853–1858 (1996).
- Legendre, P. The glycinergic inhibitory synapse. *Cell. Mol. Life Sci.* **58**, 760–793 (2001).
- Lynch, J. W. & Callister, R. J. Glycine receptors: a new therapeutic target in pain pathways. *Curr. Opin. Investig. Drugs* **7**, 48–53 (2006).
- Bode, A. & Lynch, J. W. The impact of human hyperekplexia mutations on glycine receptor structure and function. *Mol. Brain* **7**, 2 (2014).
- Young, A. B. & Snyder, S. H. Strychnine binding associated with glycine receptors of the central nervous system. *Proc. Natl Acad. Sci. USA* **70**, 2832–2836 (1973).
- Pfeiffer, F., Graham, D. & Betz, H. Purification by affinity chromatography of the glycine receptor of rat spinal cord. *J. Biol. Chem.* **257**, 9389–9393 (1982).
- Graham, D., Pfeiffer, F. & Betz, H. Photoaffinity-labelling of the glycine receptor of rat spinal cord. *Eur. J. Biochem.* **131**, 519–525 (1983).
- Bormann, J., Hamill, O. P. & Sakmann, B. Mechanism of anion permeation through channels gated by glycine and gamma-aminobutyric acid in mouse cultured spinal neurones. *J. Physiol.* **385**, 243–286 (1987).
- Fatima-Shad, K. & Barry, P. H. Anion permeation in GABA- and glycine-gated channels of mammalian cultured hippocampal neurons. *Proc. R. Soc. Lond. B* **253**, 69–75 (1993).
- Shan, Q., Hadrill, J. L. & Lynch, J. W. Ivermectin, an unconventional agonist of the glycine receptor chloride channel. *J. Biol. Chem.* **276**, 12556–12564 (2001).
- Changeux, J. P. & Edelstein, S. J. Allosteric receptors after 30 years. *Neuron* **21**, 959–980 (1998).
- Miller, P. S. & Smart, T. G. Binding, activation and modulation of Cys-loop receptors. *Trends Pharmacol. Sci.* **31**, 161–174 (2010).
- Bocquet, N. et al. X-ray structure of a pentameric ligand-gated ion channel in an apparently open conformation. *Nature* **457**, 111–114 (2009).
- Hilf, R. J. & Dutzler, R. Structure of a potentially open state of a proton-activated pentameric ligand-gated ion channel. *Nature* **457**, 115–118 (2009).
- Spurny, R. et al. Pentameric ligand-gated ion channel ELIC is activated by GABA and modulated by benzodiazepines. *Proc. Natl Acad. Sci. USA* **109**, E3028–E3034 (2012).
- Sauguet, L. et al. Crystal structures of a pentameric ligand-gated ion channel provide a mechanism for activation. *Proc. Natl Acad. Sci. USA* **111**, 966–971 (2014).
- Moraga-Cid, G. et al. Allosteric and hyperekplexic mutant phenotypes investigated on an $\alpha 1$ glycine receptor transmembrane structure. *Proc. Natl Acad. Sci. USA* **112**, 2865–2870 (2015).
- Miyazawa, A., Fujiyoshi, Y. & Unwin, N. Structure and gating mechanism of the acetylcholine receptor pore. *Nature* **423**, 949–955 (2003).
- Unwin, N. Refined structure of the nicotinic acetylcholine receptor. *J. Mol. Biol.* **346**, 967–989 (2005).
- Hibbs, R. E. & Gouaux, E. Principles of activation and permeation in an anion-selective Cys-loop receptor. *Nature* **474**, 54–60 (2011).
- Althoff, T., Hibbs, R. E., Banerjee, S. & Gouaux, E. X-ray structures of GluCl in apo states reveal a gating mechanism of Cys-loop receptors. *Nature* **512**, 333–337 (2014).
- Hassaine, G. et al. X-ray structure of the mouse serotonin 5-HT₃ receptor. *Nature* **512**, 276–281 (2014).
- Miller, P. S. & Aricescu, A. R. Crystal structure of a human GABA_A receptor. *Nature* **512**, 270–275 (2014).
- daCosta, C. J. & Baenziger, J. E. Gating of pentameric ligand-gated ion channels: structural insights and ambiguities. *Structure* **21**, 1271–1283 (2013).
- Cecchini, M. & Changeux, J. P. The nicotinic acetylcholine receptor and its prokaryotic homologues: structure, conformational transitions & allosteric modulation. *Neuropharmacology* **96**, 137–149 (2015).
- Hille, B. *Ion Channels of Excitable Membranes* (Sinauer Associates, 2001).
- Kinde, M. N. et al. Conformational changes underlying desensitization of the pentameric ligand-gated ion channel ELIC. *Structure* **23**, 995–1004 (2015).
- Gielen, M., Thomas, P. & Smart, T. G. The desensitization gate of inhibitory Cys-loop receptors. *Nature Commun.* **6**, 6829 (2015).
- Akabas, M. H. Using molecular dynamics to elucidate the structural basis for function in pLGICs. *Proc. Natl Acad. Sci. USA* **110**, 16700–16701 (2013).
- Brams, M. et al. A structural and mutagenic blueprint for molecular recognition of strychnine and *d*-tubocurarine by different cys-loop receptors. *PLoS Biol.* **9**, e1001034 (2011).
- Vandenberg, R. J., Handford, C. A. & Schofield, P. R. Distinct agonist- and antagonist-binding sites on the glycine receptor. *Neuron* **9**, 491–496 (1992).
- Marvizón, J. C. et al. The glycine receptor: pharmacological studies and mathematical modeling of the allosteric interaction between the glycine- and strychnine-binding sites. *Mol. Pharmacol.* **30**, 590–597 (1986).
- Ruiz-Gómez, A., Morato, E., García-Calvo, M., Valdivieso, F. & Mayor, F. Jr. Localization of the strychnine binding site on the 48-kilodalton subunit of the glycine receptor. *Biochemistry* **29**, 7033–7040 (1990).
- Rajendra, S. & Schofield, P. R. Molecular mechanisms of inherited startle syndromes. *Trends Neurosci.* **18**, 80–82 (1995).
- Rajendra, S. et al. The unique extracellular disulfide loop of the glycine receptor is a principal ligand binding element. *EMBO J.* **14**, 2987–2998 (1995).
- Yu, R. et al. Agonist and antagonist binding in human glycine receptors. *Biochemistry* **53**, 6041–6051 (2014).
- Grudzinska, J. et al. The β subunit determines the ligand binding properties of synaptic glycine receptors. *Neuron* **45**, 727–739 (2005).
- Mukhtasimova, N., Free, C. & Sine, S. M. Initial coupling of binding to gating mediated by conserved residues in the muscle nicotinic receptor. *J. Gen. Physiol.* **126**, 23–39 (2005).
- Hansen, S. B. et al. Structures of Aplysia AChBP complexes with nicotinic agonists and antagonists reveal distinctive binding interfaces and conformations. *EMBO J.* **24**, 3635–3646 (2005).
- Purohit, P. & Auerbach, A. Loop C and the mechanism of acetylcholine receptor-channel gating. *J. Gen. Physiol.* **141**, 467–478 (2013).
- Celie, P. H. et al. Nicotine and carbamylcholine binding to nicotinic acetylcholine receptors as studied in AChBP crystal structures. *Neuron* **41**, 907–914 (2004).
- Huang, S. et al. Complex between α -bungarotoxin and an $\alpha 7$ nicotinic receptor ligand-binding domain chimera. *Biochem. J.* **454**, 303–310 (2013).
- Lynagh, T. & Lynch, J. W. An improved ivermectin-activated chloride channel receptor for inhibiting electrical activity in defined neuronal populations. *J. Biol. Chem.* **285**, 14890–14897 (2010).
- Purohit, P., Gupta, S., Jaday, S. & Auerbach, A. Functional anatomy of an allosteric protein. *Nature Commun.* **4**, 2984 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We are grateful to Z. H. Yu, N. Grigorieff, J. Cruz and C. Hong (Janelia Campus), C. Arthur (FEI) and M. Braunfeld (UCSF) for assistance with microscope operation, data collection and for comments, and to R. Stites, M. Hakanson and A. Trzynka (OHSU) for computational support. We acknowledge the support of R. Goodman and J. Gray. Microscopy at Oregon Health & Science University (OHSU) was performed at the Multiscale Microscopy Core (MMC) with technical support from the OHSU-FEI Living Lab, Intel and the OHSU Center for Spatial Systems Biomedicine (OCSSB). We thank L. Vaskalis for help with illustrations and H. Owen for proofreading. R. Hibbs is gratefully acknowledged for pre-screening the GlyR constructs and D. P. Claxton for optimizing the constructs. We thank Gouaux and Bacongus laboratory members for discussions. This work was supported by the National Institute of Health (E.G.). E.G. is an investigator with the Howard Hughes Medical Institute.

Author Contributions J.D., W.L. and E.G. designed the project, J.D. and W.L. performed sample preparation, cryo-EM data collection and data analysis, J.D., W.L. and E.G. wrote the manuscript, S.W. and Y.C. assisted in cryo-EM experiments at UCSF and participated in discussion and editing of the manuscript.

Author Information Three three-dimensional cryo-EM density maps and coordinates of $\alpha 1$ glycine receptors in strychnine-bound, glycine-bound and glycine/ivermectin-bound forms have been deposited in the Electron Microscopy Data Bank under the accession numbers EMD-6344, EMD-6345, and EMD-6346 and deposited in the RCSB Protein Data Bank under the accession codes 3JAD, 3JAE, and 3JAF. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to E.G. (gouaux@ohsu.edu).

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Receptor constructs. The gene encoding the zebrafish $\alpha 1$ glycine receptor (NP_571477) construct shares 92% amino acid similarity with the human $\alpha 1$ GlyR (Supplementary Fig. 1). The construct was chemically synthesized and further modified in accordance with the GluCl_{crist} construct, which involved removing 8 residues (Arg26–Pro33) from the amino terminus and 10 residues (Ile435–Gln444) from the carboxyl terminus, and replacing the long, highly variable M3/M4 loop (Gln334–Arg400) with the Ala–Gly–Thr tripeptide. This construct is named GlyR_{EM} and was subcloned into the pFastBac1 vector for baculovirus expression in Sf9 insect cells. A thrombin cleavage site (Leu–Val–Pro–Arg–Ser) and an octa-histidine tag were introduced at the carboxyl terminus of GlyR_{EM}.

Expression and purification. The bacmid and baculovirus of GlyR_{EM} in pFastBac1 were generated using standard methods²⁶. GlyR_{EM} P2 virus was used to infect Sf9 insect cells at 27 °C. At 72 h post-infection, the cells were collected and disrupted by sonication. The cell debris was removed by centrifugation at 8,000 r.p.m. (1,816g) and membranes were pelleted from the supernatant by centrifugation for 1 h at 40K (Ti45 rotor) in 150 mM NaCl, 20 mM Tris 8.0 (TBS buffer) in the presence of 1 mM PMSF, 0.8 μ M aprotinin, 2 μ g ml^{−1} leupeptin, and 2 mM pepstatin A. The receptor was extracted from membranes with a buffer (affinity buffer) containing 1 mM *n*-dodecyl- β -D-maltopyranoside (C₁₂M), 200 mM NaCl and 20 mM Tris 8.0 for 2 h at 4 °C. The solubilized materials were incubated with TALON resin overnight and washed with affinity buffer in presence of 35 mM imidazole. The receptors were eluted with 250 mM imidazole at pH 8.0 and were concentrated for overnight digestion by 1:100 (w/w) thrombin at 4 °C. The receptors were further purified by size-exclusion chromatography (SEC) in a buffer containing 150 mM NaCl, 20 mM Tris 8.0 and 1 mM C₁₂M (SEC buffer) at 4 °C. Fractions containing the receptor were pooled and concentrated to 3.3 mg ml^{−1}.

Two-electrode voltage clamp electrophysiology (TEVC). The zebrafish $\alpha 1$ GlyR_{EM} was subcloned into the pGEM vector for TEVC experiments. The RNAs were then transcribed using the mMessage mMachine T7 Ultra Kit (Ambion). *Xenopus laevis* oocytes were injected with 30 ng of mRNA and were incubated at 18 °C for 2–3 days in a solution containing 96 mM NaCl, 2 mM KCl, 1 mM MgCl₂, 1.8 mM CaCl₂, 5 mM HEPES pH 7.5, and 250 μ g ml^{−1} amikacin (incubation buffer). Borosilicate pipettes were filled with 3 M KCl. The stock solutions of strychnine and glycine were prepared in H₂O at concentrations of 50 mM and 2 M, while ivermectin and picrotoxinin were solved in dimethyl sulfoxide (DMSO) at concentrations of 100 mM and 1 M, respectively. The ligands were solved in recording solution containing 96 mM NaCl, 2 mM KCl, 1 mM MgCl₂, 1.8 mM CaCl₂, and 5 mM HEPES pH 7.5. Recordings were performed at −60 mV. All recording experiments were performed three times independently. Half-maximal concentration of glycine (EC₅₀) and Hill coefficient (n_H) values were obtained using the Hill equation fitted with a nonlinear least squares algorithm using GraphPad Prism. The result was averaged from three independent experiments and the error bars represent s.e.m.

Ligand-binding assays. The strychnine binding constant was determined by scintillation-proximity assay (SPA). Purified GlyR_{EM}–His₈ (20 nM) was incubated with 1 mg ml^{−1} copper yttrium silicate (Cu–YSi) beads (Perkin Elmer) and ³H-labelled strychnine (1:9 ³H:¹H) in SEC buffer with a final volume of 100 μ l. Non-specific binding was determined by the addition of 100 mM imidazole. Assay plates were read using a MicroBeta TriLux 1450 LSC and luminescence counter and data were fit to the Hill equation using GraphPad Prism.

Crystallization and molecular replacement. GlyR_{EM} was crystallized using the hanging-drop vapour-diffusion method in the presence of 2 mM glycine. One microlitre of protein solution (1.8 mg ml^{−1}) was mixed with 1 μ l reservoir solution containing 30% PEG400, 0.1 M MES 6.5, and 0.2 M CaCl₂. Diffraction data was collected on the 8.2.1 beamline at the Advanced Light Source of Lawrence Berkeley Laboratories (ALS), using a wavelength of 0.9762 Å. The data were indexed, integrated and scaled using XDS⁵¹. The space group is P₂₁2₁2₁ and cell parameters are $a = 117.77$ Å, $b = 121.35$ Å, $c = 503.81$ Å and $\alpha = \beta = \gamma = 90^\circ$. Each asymmetric unit contains two pentameric receptors (A and B). Due to radiation damage, the overall completeness is only 62.9% (62.9%) to a resolution of 4.32 Å. R_{meas} and $CC_{1/2}$ are 0.101 (0.529) and 0.998 (0.309), respectively. Values in parentheses represent the highest resolution shell (4.46–4.35 Å). The structure was solved by molecular replacement using PHASER, with the crystal structure of GluCl (PDB code: 3RHW) as the initial search model. The LLG and TFZ of the solution are 615 and 20.6, respectively. The overall map quality of the pentamer A is better than that of the pentamer B. In pentamer A, the electron density is best for the transmembrane helices, where grooves and ridges can be

observed. In contrast, the densities for extracellular domains are partially missing, likely caused by a lack of crystal contacts.

Sample preparation and data acquisition for cryo-EM analysis. Purified GlyR_{EM} in C₁₂M was mixed with 10 mM glycine/5 μ M ivermectin, with 1 mM strychnine or with 10 mM glycine a few hours before grid preparation. Next, 2.5 μ l of protein sample at a concentration of 3.3 mg ml^{−1} was applied to a glow-discharged (10 s on each side) Quantifoil holey carbon grid (copper, 1.2 μ M/1.3 μ M hole size/hole space, 200 mesh), blotted using a Vitrobot Mark III (FEI company) using 3.5 s blotting time with 100% humidity, and then plunge-frozen in liquid ethane cooled by liquid nitrogen.

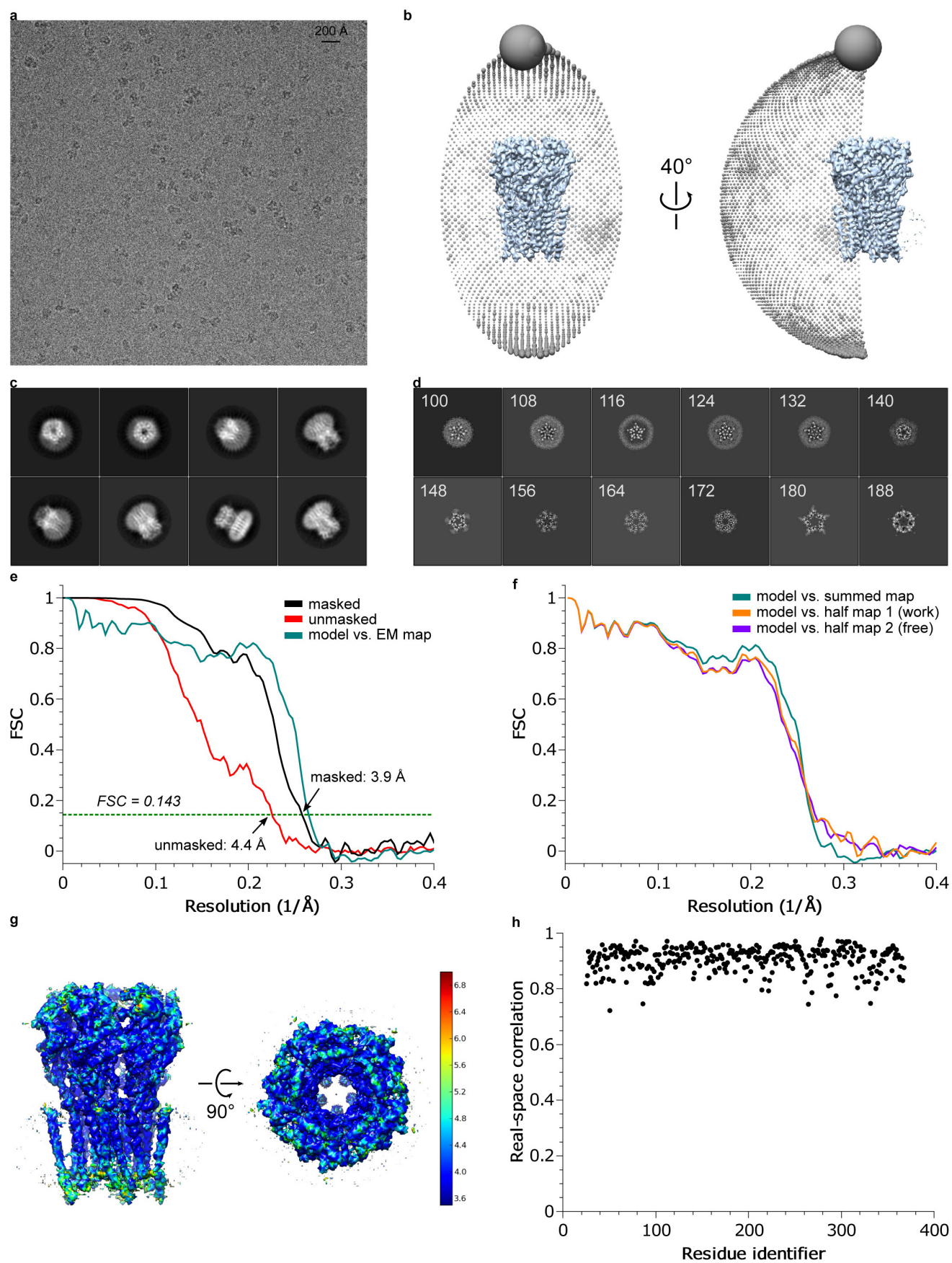
The data sets were collected on a Titan Krios cryo-electron microscope (FEI company) at the CryoEM Facility at the Janelia Research Campus or on a Polara microscope (FEI) at UCSF. The data for the glycine/ivermectin-bound state was collected on Titan Krios I and the data for the strychnine-bound state was collected on Titan Krios II at Janelia. Krios I is equipped with a CETCOR Image Corrector for spherical aberration correction and a Gatan Image Filter (GIF). A 30 eV energy slit as well as a 70 μ M objective aperture (corresponding to a cutoff of 2 Å) was used during data collection. The Image Corrector tuned the C_s from an original 2.7 to 0.01 mm. The data for the glycine-bound form was collected on the TF30 Polara at UCSF. All the microscopes are equipped with a field emission source and operated at 300 kV. Images were recorded on the Gatan K2 Summit direct electron detector operated in super-resolution counting mode. At Janelia the dose rate was 10e[−] per pixel per s, determined in ‘empty’ holes. At UCSF the dose rate was 10.9e[−] per pixel per s, determined through a layer of vitreous ice. The total exposure time was either 5 s or 6 s, with an accumulation time of 0.2 s for each frame. The dose-fractionated images were recorded using the automated acquisition program SerialEM or the semi-automated acquisition program UCSFImage4 (written by X. Li) at Janelia or UCSF, respectively. Nominal defocus values ranged from −1.5 to −2.5 μ M or −1.5 to −3.0 μ M (Extended Data Table 1).

Image processing. Super-resolution counting images were 2 × 2 binned in Fourier space, resulting in a pixel size of 1.0400 Å (Krios I), 1.0100 Å (Krios II), or 1.2156 Å (Polara). Motion correction was done using MotionCorr⁵², with a *B*-factor of 250 or 1,000 pixels squared. All the motion-corrected frames (25 or 30) were summed to a single micrograph for subsequent processing with RELION⁵³. Defocus values were estimated using CTFFIND3 (ref. 54). For each data set, approximately 1,000 particles were manually picked for an initial reference-free 2D classification. Seven to eight representative 2D class averages were selected as templates for automated particle picking. The auto-picked particles were visually checked and false positives were removed. The particles were further cleaned-up by two rounds of 2D classification using RELION. An initial 3D model was generated from the GluCl crystal structure²⁶ (PDB code: 3RHW) and low-pass filtered to 50 Å using EMAN2 (ref. 55). No symmetry was applied during 3D classification in RELION. Classes with characteristic features of Cys-loop receptors were then selected. Particles belonging to the chosen classes were used for 3D auto-refinement using RELION, followed by 3D refinement with movie particles, in which the movement of each individual particle was determined. In the subsequent ‘polishing’ step, the movement of each particle was corrected. In addition, resolution-dependent radiation damage was modelled by performing a weighted average of all aligned movie frames for each particle, using a *B*-factor estimated by the polishing algorithm. The resulting ‘shiny’ particles were used for a final round of 3D classification and refinement. The polishing procedure typically increased the estimated resolution by 0.2–0.3 Å. Five-fold symmetry was applied during all refinement steps. In the post-processing step in RELION, a soft mask was calculated and applied to the two half-maps before the Fourier shell coefficient (FSC) was calculated. *B*-factor estimation and map sharpening were also performed in the post-processing step. The resolutions reported in Extended Data Table 1 are based on the gold standard FSC 0.143 criteria⁵⁶. Further details related to data processing are summarized in Extended Data Table 1.

Model building. A molecular replacement solution from a low-resolution and incomplete X-ray diffraction data, using the GluCl structure (PDB code: 3RHW) as a search probe, was used as an initial model (see the crystallization and molecular replacement section). The mismatched residues are replaced with the correct ones using the SWISS-MODEL online server⁵⁷. The resulting model was fit into the density map of the glycine/ivermectin-bound state using UCSF Chimera⁵⁸. Further model building and real space refinement was done using COOT⁵⁹. For the other GlyR structures, the glycine/ivermectin-bound structure was used as a starting model. Strict five-fold symmetry was used throughout model building and refinement. The final masked maps were put into P1 unit cells with cell dimensions equal to the box sizes used to extract particles (Extended Data Table 1) and structure factors were calculated using the program SFTOOLS in the CCP4 suite⁶⁰. The resulting structure factors were used for maximum

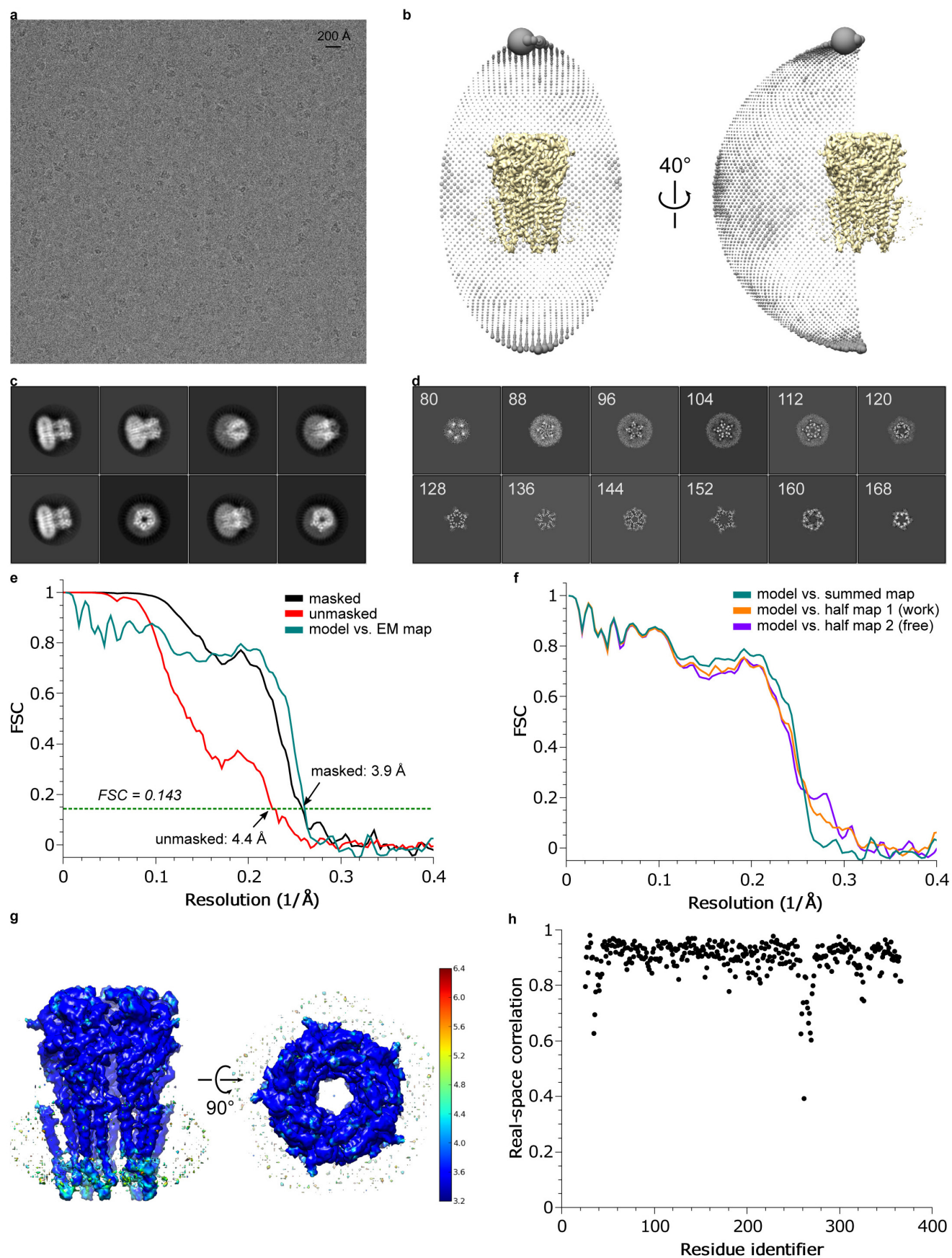
likelihood refinement in reciprocal space using Phenix.refine with secondary structure and NCS restraints. For cross validation⁶¹, the refined structures were randomly displaced by 0.1 Å. The displaced models were then refined against one of the half maps produced by RELION following the same procedure described above. FSC curves were calculated between the refined model and half map 1 ('work', used for refinement), refined model and half map 2 ('free', not used for refinement), and the refined model and summed map. No significant separation of work and free FSC curves was observed, suggesting the atomic models were not over-refined. The geometries of the atomic models were evaluated using MolProbity⁶². Our models have the same helical registers as the crystal structures of GluCl and the GABA_A and 5-HT₃ receptors, but differ by one turn from that of the 4 Å cryo-EM Torpedo AChR structure. All figures were prepared using Pymol (Schrödinger)⁶³, UCSF Chimera and Prism 5 (GraphPad, La Jolla). Pore radii were calculated using the program HOLE⁶⁴.

51. Kabsch, W. XDS. *Acta Crystallogr. D* **66**, 125–132 (2010).
52. Li, X. *et al.* Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nature Methods* **10**, 584–590 (2013).
53. Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).
54. Mindell, J. A. & Grigorieff, N. Accurate determination of local defocus and specimen tilt in electron microscopy. *J. Struct. Biol.* **142**, 334–347 (2003).
55. Tang, G. *et al.* EMAN2: an extensible image processing suite for electron microscopy. *J. Struct. Biol.* **157**, 38–46 (2007).
56. Scheres, S. H. & Chen, S. Prevention of overfitting in cryo-EM structure determination. *Nature Methods* **9**, 853–854 (2012).
57. Arnold, K., Bordoli, L., Kopp, J. & Schwede, T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* **22**, 195–201 (2006).
58. Pettersen, E. F. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
59. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
60. Winn, M. D. *et al.* Overview of the CCP4 suite and current developments. *Acta Crystallogr. D* **67**, 235–242 (2011).
61. Amunts, A. *et al.* Structure of the yeast mitochondrial large ribosomal subunit. *Science* **343**, 1485–1489 (2014).
62. Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66**, 12–21 (2010).
63. DeLano, W. L. The PyMOL Molecular Graphics System. (DeLano Scientific, San Carlos, USA, 2002).
64. Smart, O. S., Neduevilil, J. G., Wang, X., Wallace, B. A. & Sansom, M. S. HOLE: a program for the analysis of the pore dimensions of ion channel structural models. *J. Mol. Graph.* **14**, 354–360 (1996).
65. Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nature Methods* **11**, 63–65 (2014).

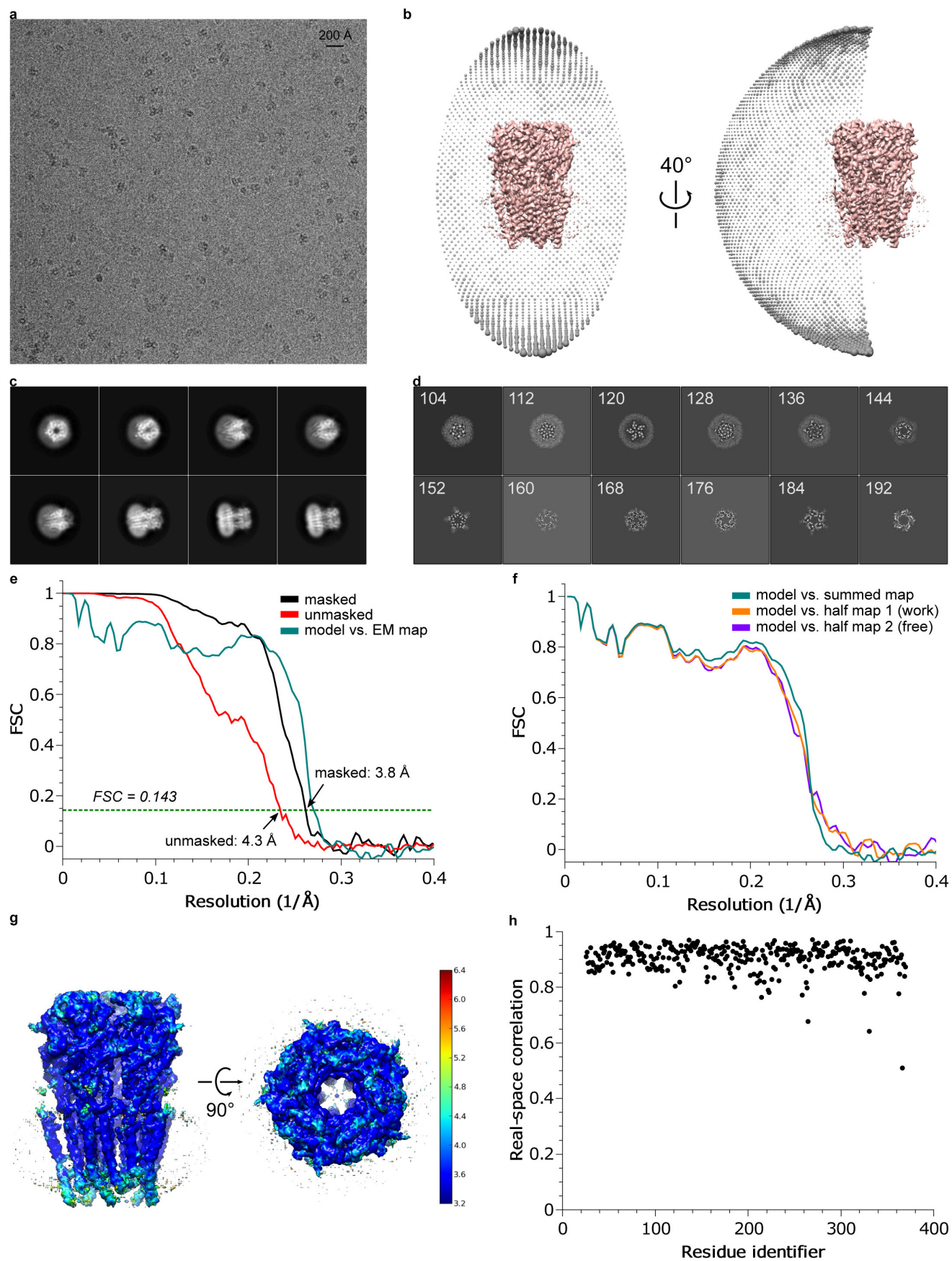


Extended Data Figure 1 | Three-dimensional reconstruction of strychnine-bound GlyR. **a**, A representative micrograph (out of 1,829 micrographs) of strychnine-bound GlyR in vitreous ice. **b**, **c**, Angular distribution of particle projections (**b**), and selected 2D classes (**c**) are shown. In **c**, the radius of the sphere is proportional to the number of particles assigned to it. The plot is drawn with respect to the 3D reconstruction shown in the centre, taking the C5 symmetry of the receptor into account. **d**, Selected 'slice' views of the final reconstruction along the pore axis. The slice numbers are indicated, starting from the intracellular side. **e**, FSC curves for the density maps before (red) and

after (black) post-processing in RELION. The FSC curve between the refined atomic model and the final reconstruction map is shown in green. **f**, FSC curves for cross-validation: model versus summed map (full data set, green), model versus half map 1 (used in test refinement, orange) and model versus half map 2 (not used in test refinement, purple). **g**, Unfiltered and unsharpened 3D density map coloured according to local resolution estimated using RESMAP⁶⁵. **h**, Real-space correlation between atomic model and density map calculated using PHENIX.

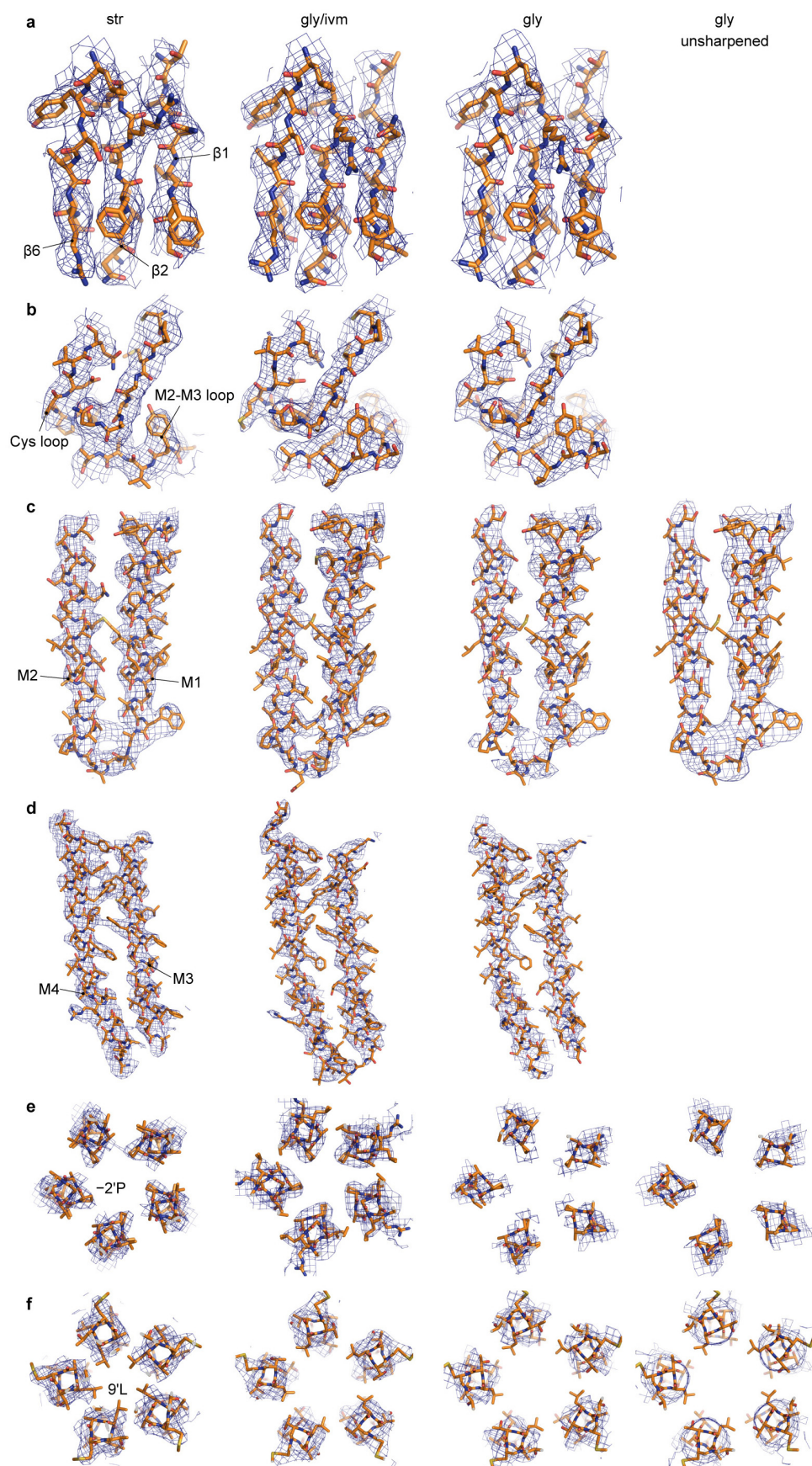


Extended Data Figure 2 | 3D reconstruction of glycine-bound GlyR. **a**, A representative micrograph (out of 1,460 micrographs) of glycine-bound GlyR in vitreous ice. **b**, **c**, Angular distribution of particle projections (**b**) and selected 2D classes (**c**) are shown. In **c**, the radius of the sphere is proportional to the number of particles assigned to it. The plot is drawn with respect to the 3D reconstruction shown in the centre, taking the C5 symmetry of the receptor into account. **d**, Selected 'slice' views of the final reconstruction along the pore axis. The slice numbers are indicated, starting from the intracellular side. **e**, FSC curves for the density maps before (red) and after (black) post-processing in RELION. The FSC curve between the refined atomic model and the final reconstruction map is shown in green. **f**, FSC curves for cross-validation: model versus summed map (full data set, green), model versus half map 1 (used in test refinement, orange) and model versus half map 2 (not used in test refinement, purple). **g**, Unfiltered and unsharpened 3D density map coloured according to local resolution estimated using RESMAP. **h**, Real-space correlation between atomic model and density map calculated using PHENIX.



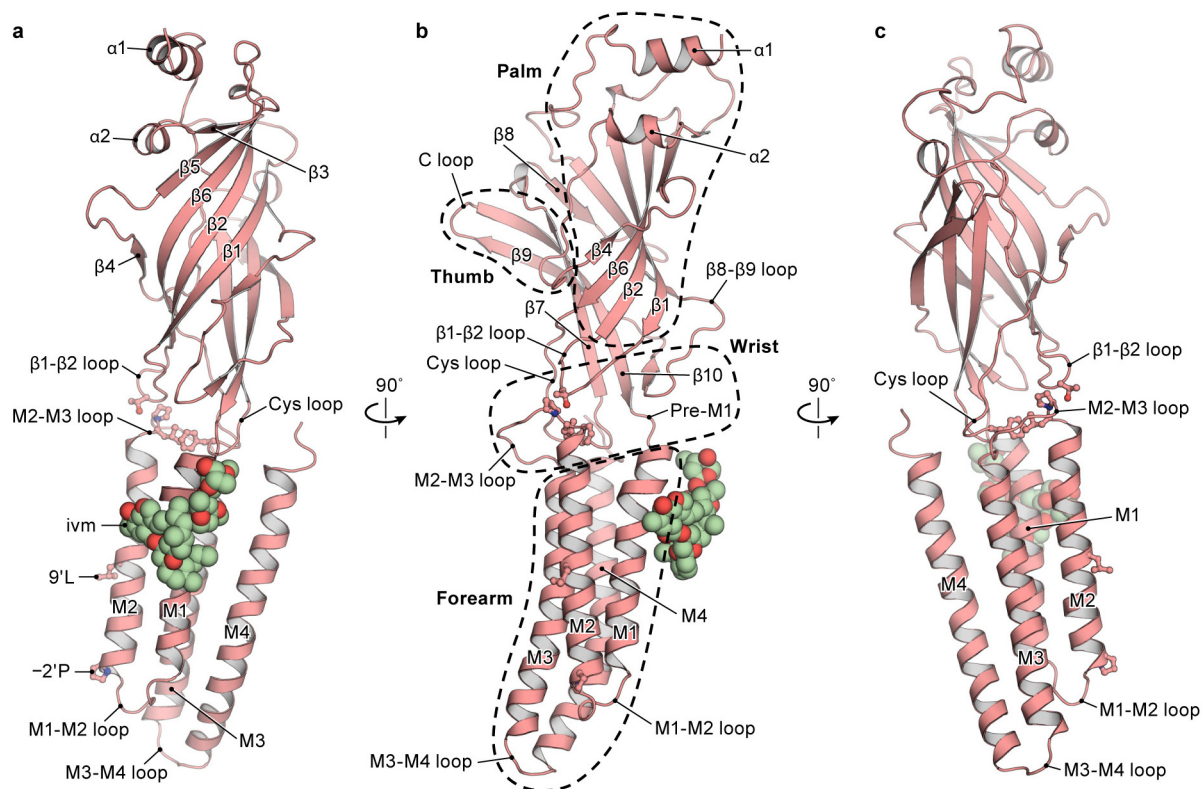
Extended Data Figure 3 | 3D reconstruction of glycine/ivermectin-bound GlyR. **a**, A representative micrograph (out of 2,489 micrographs) of glycine/ivermectin-bound GlyR in vitreous ice. **b**, **c**, Angular distribution of particle projections (**b**) and selected 2D classes (**c**) are shown. **c**, The radius of the sphere is proportional to the number of particles assigned to it. The plot is drawn with respect to the 3D reconstruction shown in the centre, taking the C5 symmetry of the receptor into account. **d**, Selected 'slice' views of the final reconstruction along the pore axis. The slice numbers are indicated, starting from the intracellular side. **e**, FSC curves for the density maps before (red) and

after (black) post-processing in RELION. The FSC curve between the refined atomic model and the final reconstruction map is shown in green. **f**, FSC curves for cross-validation: model versus summed map (full data set, green), model versus half map 1 (used in test refinement, orange) and model versus half map 2 (not used in test refinement, purple). **g**, Unfiltered and unsharpened 3D density map coloured according to local resolution estimated using RESMAP. **h**, Real-space correlation between atomic model and density map calculated using PHENIX.



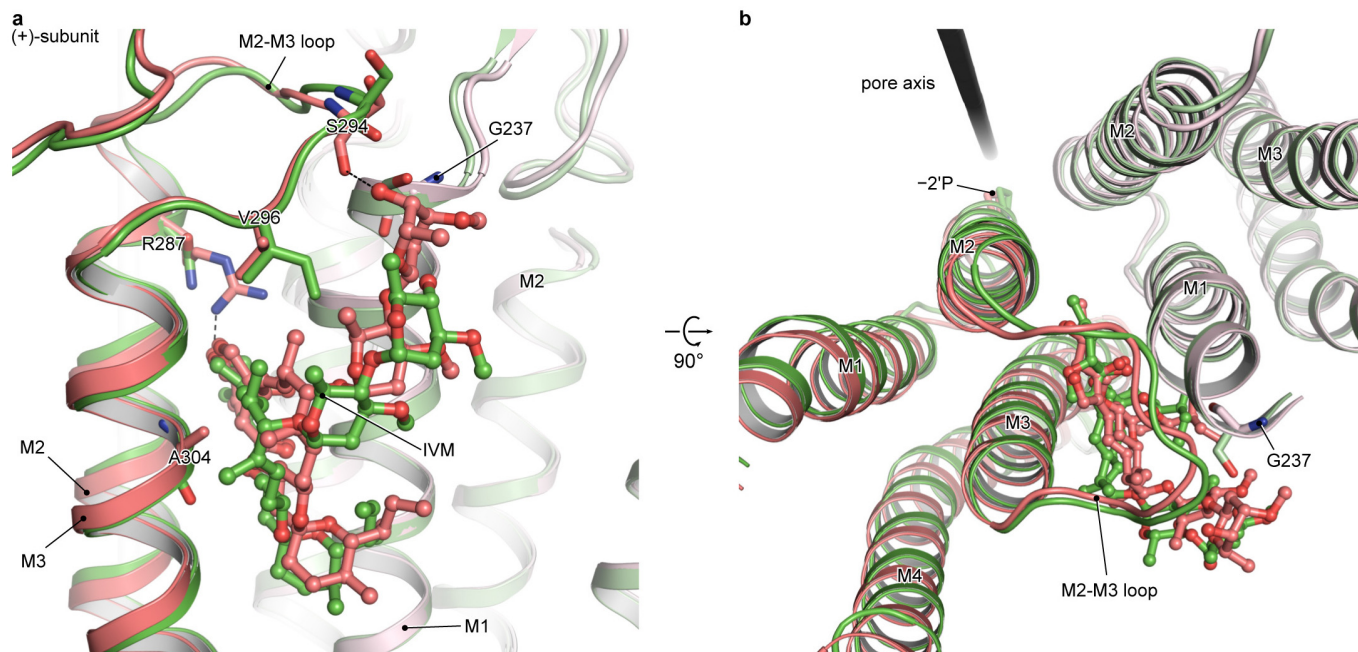
Extended Data Figure 4 | Representative densities of the three reconstructions of GlyR. Densities are sharpened using RELION unless indicated otherwise. The densities in each panel are for the strychnine-, glycine/ivermectin-, glycine-, and unsharpened glycine-bound states, respectively, from left to right. **a**, Representative densities of the β -sheets in

ECD, contoured at 8σ . **b**, Densities of Cys loop and the M2–M3 loop, contoured at 7σ . **c**, Densities of helices M1 and M2, contoured at 7σ . **d**, Densities of M3 and M4, contoured at 7σ . **e**, Densities of –2'Pro, contoured at 7σ except for the glycine-bound state (6.5σ). **f**, Densities of 9'Leu, contoured at 6.0σ except for the glycine-bound state (5.0σ).



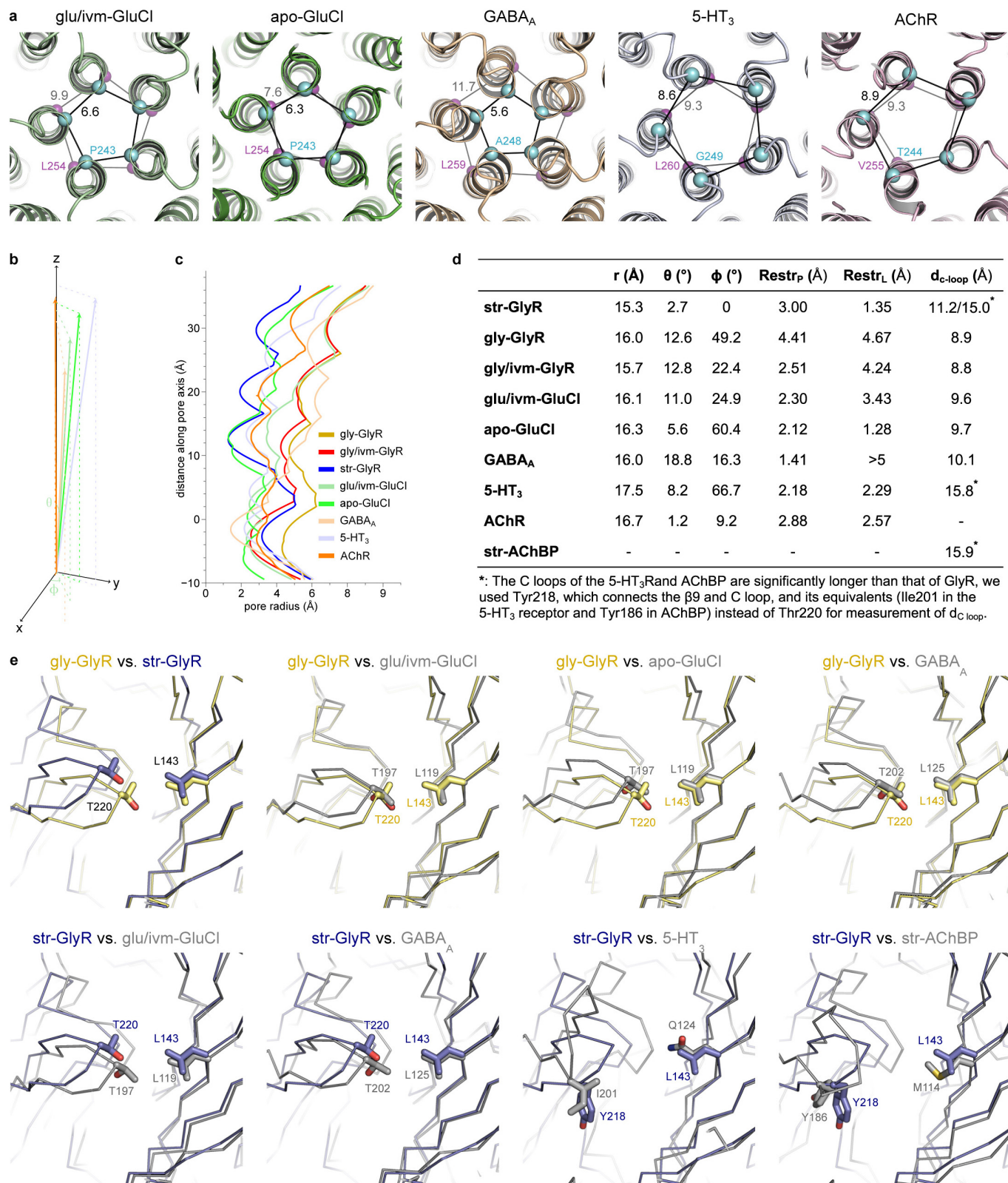
Extended Data Figure 5 | A single subunit of glycine/ivermectin bound GlyR. **a–c,** Viewed in parallel to the membrane plane, with secondary structure elements labelled. **b,** The domain arrangement resembles an upright forearm,

clad with a mitten, consisting of thumb (C loop), palm (β -strands of ECD) and ligand (ECD–TMD interface).



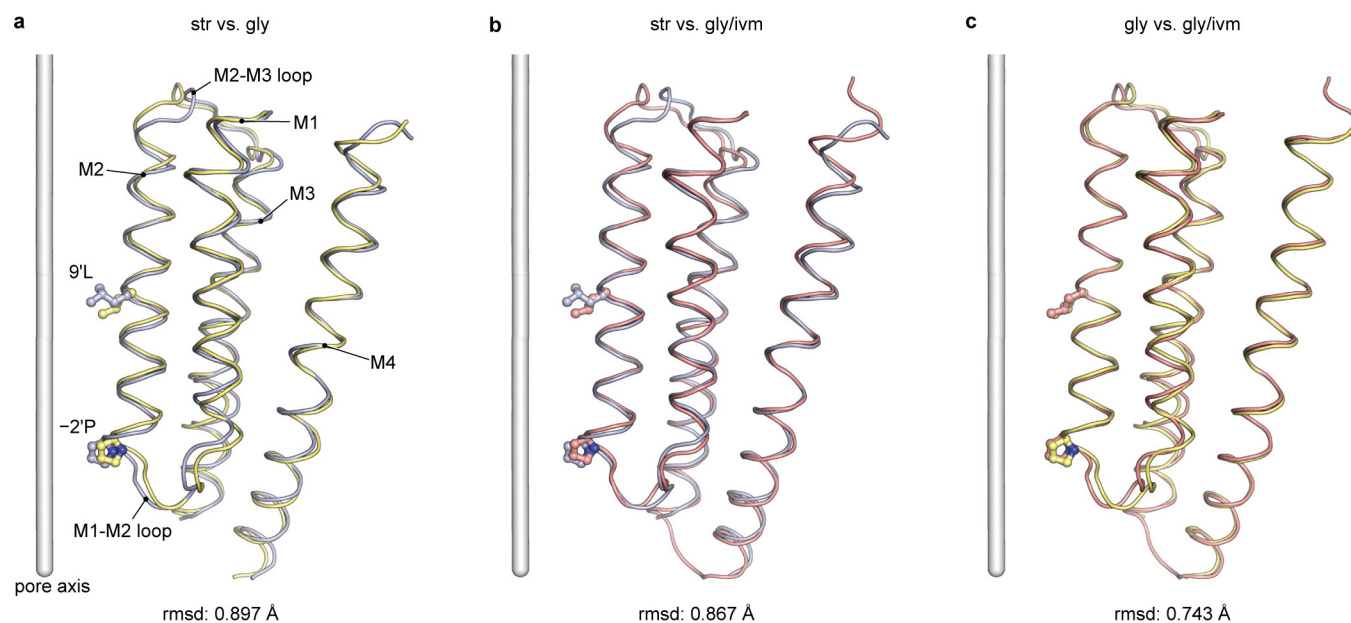
Extended Data Figure 6 | Comparison of ivermectin-binding site in GlyR (red) and GluCl (green). **a**, Viewed in parallel to the membrane. **b**, Viewed from the extracellular side. The (+)-subunits are shown in darker colours. The residue corresponding to Arg287, which forms a hydrogen bond with the ivermectin in GlyR, is an asparagine (Asn264) in GluCl. The corresponding residue of Val296 in the M2–M3 loop of GlyR is an isoleucine (Ile273) in GluCl,

whose larger side chain prevents the upper tip of ivermectin from approaching and interacting with the main chain oxygen atom of Ser721 in the M2–M3 loop (Ser294 in GlyR). The Gly237 in the M1 and Ala304 in the M3 of GlyR are Ser217 and Gly281 in GluCl, respectively. Such differences on side chains weaken or strengthen the interaction of ivermectin with M3 or M1 in GlyR, respectively, in comparison to that in GluCl.



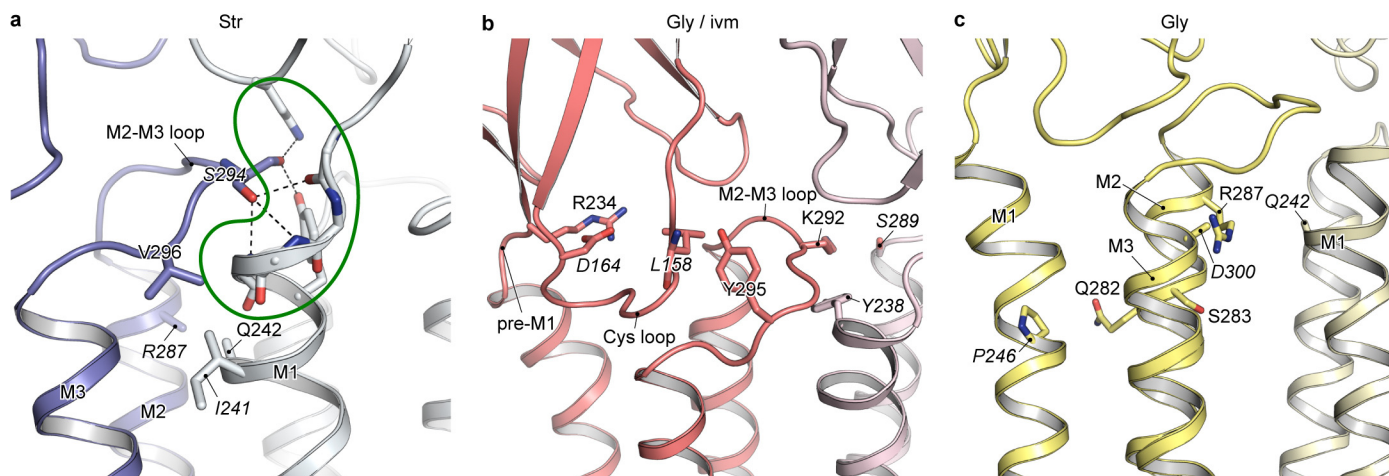
Extended Data Figure 7 | Comparison of GlyR with other Cys-loop receptors. **a**, The two restriction sites, viewed from the cytoplasmic side. The C_α of $-2'$ Pro equivalents (cyan) and $9'$ Leu equivalents (magenta) are shown as spheres. Distances between adjacent C_α atoms are labelled. **b**, Plot of the vector connecting the $-2'$ Pro C_α equivalent and $9'$ Leu C_α equivalent, with $-2'$ Pro C_α equivalent as the origin, the tilt angle θ and the rotation angle ϕ relative to the pore axis. The ϕ of strychnine-bound GlyR is arbitrarily set to zero. **c**, Pore radii as a function of distance along the pore axis, calculated using

the program HOLE, where the C_α position of $0'$ Arg is set to zero. **d**, Table showing parameters of the vector connecting the $-2'$ Pro C_α equivalent and $9'$ Leu C_α equivalent, where r is the distance from $-2'$ Pro C_α equivalent to $9'$ Leu C_α equivalent, $Restr_P$ and $Restr_L$ are the pore radii at $-2'$ Pro equivalent and $9'$ Leu equivalent, respectively. The d_{C-loop} is the distance between C_α of Thr220 equivalent and Leu143 equivalent, representing the opening of the C loop shown in **e**. **e**, Comparison of ligand-binding pockets. The side chains of marker residues are shown in stick format.



Extended Data Figure 8 | Superimposition of TMD in three GlyR structures using main chain atoms of residues Met236–Lys362. **a**, Between strychnine- and glycine–GlyR. **b**, Strychnine- and glycine/ivermectin–GlyR. **c**, Glycine- and glycine/ivermectin–GlyR. The M2–M3 loop, residues Ser289–Ala298, is excluded from the comparison. The root mean square deviations (r.m.s.d.) are

0.9, 0.9, and 0.7 Å, respectively, suggesting that the movement of the TMD is rigid-body-like. Most differences are located in the termini of transmembrane helices, which are either close to the M2–M3 loop, or close to the intracellular gate –2'Pro.



Extended Data Figure 9 | Positions of residues in which mutations are associated with human startle disease. Residues that likely interact with disease-causing residues are labelled in *italics*. **a**, The strychnine–GlyR model is used to show residues in which mutations cause spontaneous activation. The mutation of Gln242 in M1 to glutamate may enhance its electrostatic attraction to Arg287 in M2 of the adjacent subunit and tilt the upper part of M2 away from pore axis, resulting in a constitutively open channel. Alternatively, the mutation Val296Met in M2–M3 loop may cause steric collision with Ile241 in M1 of the adjacent subunit, and prevent Ser294 from interacting to the N-cap formed by pre-M1, M1 and the $\beta 8$ – $\beta 9$ loop, thereby destabilizing the closed conformation. **b**, The glycine/ivermectin–GlyR model is used to show residues in the ECD–TMD interface whose mutations reduce sensitivity to glycine and single channel conductance. The mutation of Arg234 in pre-M1

to glutamine may disturb its electrostatic interaction with Asp164 in the Cys loop. Similarly, the mutation of Tyr295 in the M2–M3 loop to cysteine or serine may disturb its interaction with the main chain nitrogen atom of Leu158 in the Cys loop. In both cases, the signal induced by agonist binding may be blocked. The mutation Lys292Glu in the M2–M3 loop possibly affects the cooperative interaction between two adjacent subunits by altering the van der Waals contacts between Lys292 and Tyr238. **c**, The glycine–GlyR model is used to show residues in M2 in which mutations reduce sensitivity to glycine and diminish single channel conductance. These mutations may directly influence the pore properties by modifying the interactions with adjacent residues, for instance, between Gln282His and Pro246, and between Arg287Gln/Leu and Gln242.

Extended Data Table 1 | Statistics of 3D reconstruction and model refinement

| | strychnine | glycine | glycine/ivermectin |
|--|-------------|-------------|--------------------|
| Data collection/processing | | | |
| Microscope | Krios | Polaris | Krios |
| Voltage (kV) | 300 | 300 | 300 |
| Defocus range (μM) | -1.5 – -3.0 | -1.5 – -2.5 | -1.5 – -3.0 |
| Exposure time (s) | 5 | 6 | 5 |
| Dose rate ($e^-/\text{pixel/s}$) | 10.0 | 10.9 | 10.0 |
| Pixel size (\AA) | 1.0100 | 1.2156 | 1.0400 |
| Particles processed | 82913 | 127276 | 160585 |
| Particles refined | 37094 | 58188 | 56957 |
| Resolution (unmasked, \AA) | 4.4 | 4.4 | 4.3 |
| Resolution (masked, \AA) | 3.9 | 3.9 | 3.8 |
| Map sharpening B-factor (\AA^2) | -146 | -151 | -156 |
| Refinement | | | |
| Cell dimensions | | | |
| $a = b = c$ (\AA) | 282.8 | 291.7 | 291.2 |
| $\alpha = \beta = \gamma$ ($^\circ$) | 90 | 90 | 90 |
| Resolution (\AA) | 3.9 | 3.9 | 3.8 |
| Number of atoms | 12705 | 12810 | 13870 |
| Protein | 12440 | 12670 | 13420 |
| Ligand | 265 | 140 | 450 |
| r.m.s. deviations | | | |
| Bond length (\AA) | 0.004 | 0.005 | 0.005 |
| Bond angle ($^\circ$) | 0.960 | 1.115 | 1.256 |
| Ramachandran plot | | | |
| Favored (%) | 97.4 | 99.1 | 99.4 |
| Allowed (%) | 2.6 | 0.9 | 0.6 |
| Disallowed (%) | 0 | 0 | 0 |

Fast-moving features in the debris disk around AU Microscopii

Anthony Boccaletti¹, Christian Thalmann², Anne-Marie Lagrange^{3,4}, Markus Janson^{5,6}, Jean-Charles Augereau^{3,4}, Glenn Schneider⁷, Julien Milli^{4,8}, Carol Grady⁹, John Debes¹⁰, Maud Langlois^{11,12}, David Mouillet^{3,4}, Thomas Henning⁶, Carsten Dominik¹³, Anne-Lise Maire¹⁴, Jean-Luc Beuzit^{3,4}, Joseph Carson^{6,15}, Kjetil Dohlen¹², Natalia Engler², Markus Feldt⁶, Thierry Fusco^{12,16}, Christian Ginski¹⁷, Julien H. Girard^{4,8}, Dean Hines¹⁰, Markus Kasper^{4,18}, Dimitri Mawet¹⁹, François Ménard²⁰, Michael R. Meyer², Claire Moutou¹², Johan Olofsson⁶, Timothy Rodigas²¹, Jean-François Sauvage^{12,16}, Joshua Schlieder^{6,22}, Hans Martin Schmid², Massimo Turatto¹⁴, Stéphane Udry²³, Farrokh Vakili²⁴, Arthur Vigan^{8,12}, Zahed Wahhaj^{8,12} & John Wisniewski²⁵

In the 1980s, excess infrared emission was discovered around main-sequence stars; subsequent direct-imaging observations revealed orbiting disks of cold dust to be the source¹. These ‘debris disks’ were thought to be by-products of planet formation because they often exhibited morphological and brightness asymmetries that may result from gravitational perturbation by planets. This was proved to be true for the β Pictoris system, in which the known planet generates an observable warp in the disk^{2–5}. The nearby, young, unusually active late-type star AU Microscopii hosts a well-studied edge-on debris disk; earlier observations in the visible and near-infrared found asymmetric localized structures in the form of intensity variations along the midplane of the disk beyond a distance of 20 astronomical units^{6–9}. Here we report high-contrast imaging that reveals a series of five large-scale features in the southeast side of the disk, at projected separations of 10–60 astronomical units, persisting over intervals of 1–4 years. All these features appear to move away from the star at projected speeds of 4–10 kilometres per second, suggesting highly eccentric or unbound trajectories if they are associated with physical entities. The origin, localization, morphology and rapid evolution of these features are difficult to reconcile with current theories.

The system AU Microscopii (AU Mic) is peculiar in many respects. The star is a flaring¹⁰ cool M1Ve type dwarf at a distance of only 9.94 ± 0.13 pc from Earth, and is a member of the β Pic Moving Group, with an age of 23 ± 3 Myr (ref. 11). Its extended (about 200 astronomical units, AU) edge-on, optically thin debris disk was first imaged at visible wavelengths from the ground¹². The current picture of the system assumes a ‘birth ring’ of gas-depleted¹⁴ planetesimals located at 35–40 AU (ref. 13). Beyond this radius, the disk is populated by small dust particles ($>0.05 \mu\text{m}$)⁹, probably driven outward by stellar wind; radiation pressure alone would be insufficient to explain the disk’s extent¹³. Following the discovery image, the system was intensively observed in 2004/2005 from the ground and space^{6–9}. Several intensity inhomogeneities in the form of clumps were reported far from the star at physical separations of 20–40 AU. Most were located

in the fainter, southeast side of the disk, while the northwest side was more uniform and approximately twice as bright. The exact positions of these structures differ slightly in the literature, possibly owing to wavelength dependencies⁹. More recently, observations obtained in August 2010 and July 2011 using the Hubble Space Telescope (HST) confirmed the presence of structures in the AU Mic debris disk¹⁵.

AU Mic was one of the prime test targets during the commissioning of SPHERE¹⁶, the planet finder instrument installed at the Very Large Telescope (VLT). It was observed on 10 August 2014, in the J band ($1.25 \mu\text{m}$) with SPHERE’s near-infrared camera IRDIS. Owing to good and stable atmospheric conditions (with seeing about $1.25''$ and wind $<10 \text{ m s}^{-1}$), the adaptive optics delivered high Strehl ratios (corresponding to 90% to 95% at the SPHERE reference wavelength $\lambda = 1.65 \mu\text{m}$), which resulted in high focal-plane contrasts of 9×10^{-5} at about $0.5''$ on average.

The disk is detected out to $7''$ (about 70 AU), as limited by the detector field of view, and as close as $0.17''$ (about 1.7 AU), below which the disk is attenuated by the coronagraph (Fig. 1). We measured a position angle (PA) of $129.5^\circ \pm 0.3^\circ$ in the southeast side. The northwest-side PA differs by $1.7^\circ \pm 0.4^\circ$ (see Methods). Although the general shape agrees with previous observations, the new SPHERE images show the morphology of the whole disk with unprecedented resolution and detail.

The most striking features revealed by the SPHERE observations are the arch- or wave-like structures close to the star in the southeastern side (annotated A to E in Fig. 1). The features A, B and C, which are located above the midplane, are closer than the ones reported earlier, and do not resemble anything previously observed in circumstellar disks. Two additional fainter structures, D and E, are observed at larger projected separations, closer to and overlapping with the midplane. In addition, they show a wavy (undulating) morphology (Fig. 1). The projected separations of these five structures span the range of ~ 10 AU to 55 AU (approximately $1.02''$, $1.70''$, $2.96''$, $4.10''$ and $5.52''$). The typical projected radial extents of the features range between approximately 5 AU (for A, the closest) to 10 AU (for E, the farthest away), and they reach elevations above the disk midplane in the range ~ 1.5 AU (for

¹LESIA, Observatoire de Paris, CNRS, Université Paris Diderot, Université Pierre et Marie Curie, 5 place Jules Janssen, 92190 Meudon, France. ²ETH Zürich, Institute for Astronomy, Wolfgang-Pauli-Strasse 27, CH-8093 Zürich, Switzerland. ³Université Grenoble Alpes, IPAG, F-38000 Grenoble, France. ⁴CNRS, IPAG, F-38000 Grenoble, France. ⁵Department of Astronomy, Stockholm University, SE-106 91 Stockholm, Sweden. ⁶Max-Planck-Institut für Astronomie, Königstuhl 17, D-69117 Heidelberg, Germany. ⁷Steward Observatory, 933 North Cherry Avenue, The University of Arizona, Tucson, Arizona 85721, USA. ⁸European Southern Observatory (ESO), Alonso de Córdova 3107, Vitacura, Casilla 19001, Santiago, Chile. ⁹Eureka Scientific, 2452 Delmer, Suite 100, Oakland, California 96602, USA. ¹⁰Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, Maryland 21218, USA. ¹¹Centre de Recherche Astrophysique de Lyon, (CNRS/ENS-L/Université Lyon 1), 9 avenue Charles André, 69561 Saint-Genis-Laval, France. ¹²Aix Marseille Université, CNRS, LAM (Laboratoire d’Astrophysique de Marseille) UMR 7326, 13388 Marseille, France. ¹³University of Amsterdam, Anton Pannekoek Institute for Astronomy, Science Park 904 1098 XH Amsterdam, The Netherlands. ¹⁴INAF-Osservatorio Astronomico di Padova, Vicolo dell’Osservatorio 5, 35122 Padova, Italy. ¹⁵Department of Physics and Astronomy, College of Charleston, South Carolina, 29424, USA. ¹⁶ONERA—The French Aerospace Laboratory, 92322 Châtillon, France. ¹⁷Sterrewacht Leiden, PO Box 9513, Niels Bohrweg 2, NL-2300RA Leiden, The Netherlands. ¹⁸European Southern Observatory (ESO), Karl Schwarzschild Strasse 2, 85748 Garching bei München, Germany. ¹⁹Department of Astronomy, California Institute of Technology, 1200 East California Boulevard, MC 249-17, Pasadena, California 91125, USA. ²⁰UMI-FCA, CNRS/INSU France (UMI 3386), and Departamento de Astronomía, Universidad de Chile, Casilla 36-D, Correo Central, Santiago, Chile. ²¹Department of Terrestrial Magnetism, Carnegie Institution of Washington, 5241 Broad Branch Road NW, Washington DC 20015, USA. ²²NASA Ames Research Center, Space Science and Astrobiology Division, MS 245-6, Moffett Field, California 94035, USA. ²³Observatoire de Genève, University of Geneva, 51 Chemin des Maillettes, 1290 Versoix, Switzerland. ²⁴Laboratoire J.-L. Lagrange, Observatoire de la Côte d’Azur (OCA), Université de Nice-Sophia Antipolis (UNS), CNRS, Campus Valrose, 06108 Nice Cedex 2, France. ²⁵Department of Physics and Astronomy, University of Oklahoma, 440 West Brooks Street, Norman, Oklahoma 73019, USA.

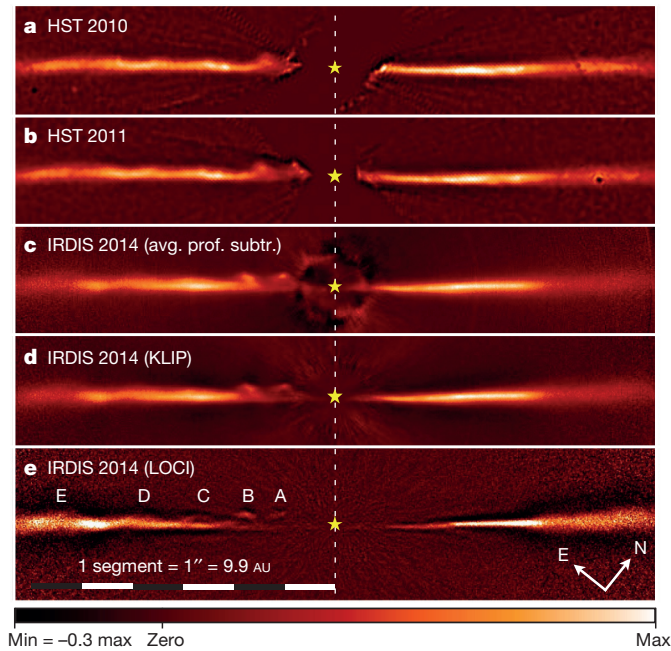


Figure 1 | High-contrast images of the AU Mic debris disk. Images are shown for the three epochs (2010.69, 2011.63 and 2014.69) at the same spatial scale; the location of AU Mic is marked with a yellow star symbol. In **a** and **b**, the HST/STIS data were processed with multi-roll point spread function (PSF)-template subtraction and unsharp mask. SPHERE/IRDIS images are displayed in **c**, **d** and **e**, for three differential imaging techniques (average profile subtraction, KLIP and LOCI) (see Methods). The intensity maps are multiplied by the square of the stellocentric distance to counteract the high dynamic range of the data and to make the disk structures A–E visible at all separations.

A) to 0.5 AU (for E). Features A and B are recovered with the visible-light instrument channel of SPHERE, as well (Methods).

To confirm the presence and reliability of these features we revisited older observations with HST's Space Telescope Imaging Spectrograph (STIS) in 2010/2011, in which a bump in the midplane was reported in the southeast side at a projected separation of about 13 AU (ref. 15). We reanalysed these data to yield separate images for the 2010 and 2011 epochs, augmented with unsharp masking to render the structures more visible. Both epochs show that this bump is equivalent to the feature B seen in the 2014 SPHERE image but situated about 4 AU closer to the star (Fig. 1), and similarly feature A is also visible from the 2011 epoch. A more careful look reveals that the HST reprocessed images also contain more features all along the midplane. Not only do the features in the SPHERE and HST images match with high fidelity across all three

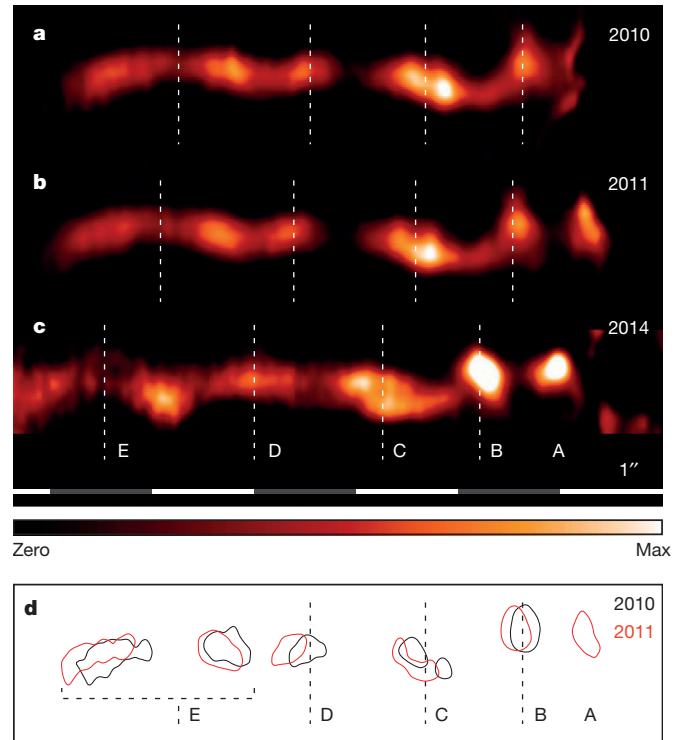


Figure 2 | Extraction of disk substructure from the southeastern side.

a–c, The images from Fig. 1a–c after unsharp masking, subtraction of the smooth main body of the disk, and stretching in the vertical direction by a factor of two (see Methods). The same persistent pattern is recovered in all three epochs, though at shifted locations, implying motion away from the star. **d**, A contour plot of the two HST epochs after more aggressive spatial filtering (Methods), which produces sharp residual features highlighting the differential motion of each feature.

epochs, but they also appear radially offset between epochs, suggesting that the features are moving away from the star, as shown in Fig. 2.

To precisely register the features we plot the disk spine's transversal excursions from the midplane and its intensity as a function of separation from the star (Fig. 3). We note that these two methods do not trace exactly the same physical structures, since the intensity maxima do not coincide with the excursion peaks for the outer features (Fig. 3a, b). Nevertheless, both methods show a persistent pattern shifting away from the star over a 4-year time frame. The five features are clearly identifiable as peaks in the excursion plot. As a general trend, the features get fainter, broader, and closer to the midplane of the disk

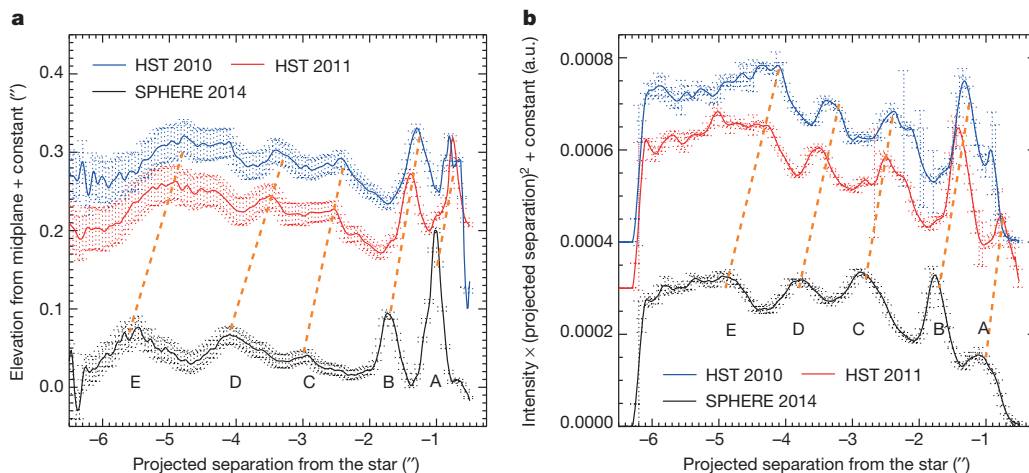


Figure 3 | Disk features across three epochs. Precise registration of the disk spine in the southeast side reveals vertical excursions (**a**) and intensity variations (multiplied by the square of the separation from the star, **b**). The SPHERE profile is an average of three data reductions (ADI, KLIP and subtraction of azimuthally averaged profile). Error bars are 1σ dispersion. The profiles are shifted vertically in proportion to the time intervals between epochs. Disk features are identified as five local maxima (A–E). Dashed orange lines roughly illustrate the possible trajectory of each feature. Feature A is undetected in 2010, being too close to the star. a.u., arbitrary units.

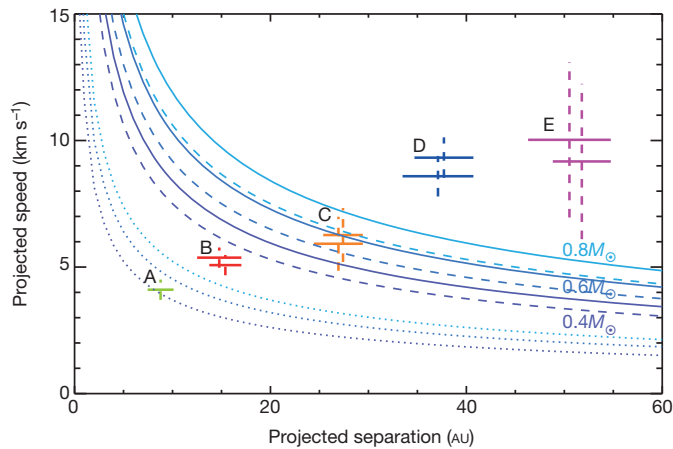


Figure 4 | Projected speeds of the disk features. The projected speeds of the five features A–E (green, red, orange, blue and magenta) are plotted against the projected separation from the star. Several orbits are shown for different mass assumptions (0.4, 0.6 and 0.8 solar masses) and two eccentricities: $e = 0$ (dotted lines), $e = 0.9$ (dashed lines). The solid lines stand for the maximum local system escape speed. Horizontal bars correspond to the range of projected separations between two epochs, while the vertical dotted lines stand for the projected speed uncertainty (peak-to-valley).

with increasing stellocentric distance (Fig. 3). Feature A is inside the blind area of the HST 2010 image. Finally, we conclude that all structures identified in 2014 are recovered in 2010 and 2011 and appear to have moved away from the star towards the southeast direction as a coherent series of patterns. The fact that the two HST epochs alone (biases being minimal) exhibit a noticeable motion is a very strong argument in favour of a real phenomenon. This motion is opposite to that of background objects given AU Mic's proper motion. The colour dependence of the grains' scattering properties cannot account for such a large displacement between the visible and the infrared.

From the three available epochs we obtained the projected speeds associated with each feature considering the excursions from the mid-plane (Fig. 4). To remain conservative the registration errors are peak-to-valley instead of 1σ dispersion. The measured speeds are in the range $4\text{--}10\text{ km s}^{-1}$. Assuming stellar mass in the range of 0.6 ± 0.2 solar masses, the projected speeds of all features beyond A are inconsistent with circular orbits. The speeds of features B and C are compatible with elliptical orbits, but require minimum eccentricities of ~ 0.5 and ~ 0.97 even for the high end of the stellar mass range. Features D and E are fainter and less distinct than the closer features, which makes their speed measurements less accurate. However, taking into account the error bars, D and E exceed the local system escape velocity for all stellar mass assumptions. To a lower extent, feature C has a similar behaviour for the lowest-stellar-mass assumption. If confirmed with future measurements, these speeds may indicate that at least two (and possibly three) of the features are on unbound trajectories leaving the system.

Several mechanisms were considered that might produce structures in a dusty disk, some involving a gas-rich disk, spiral waves, resonances with planetary-mass objects, stellar activity, or outflows from planets (see Methods). But the distinct morphology of the features, their high apparent speeds incompatible with low-eccentricity orbits, and their spatial localization on only one side of the disk are at odds with most scenarios. Therefore, we cannot offer a single explanation for these features; additional data are needed to do so. New HST and IRDIS imaging can monitor the morphological, photometric, and astrometric temporal evolution of the features, determine whether their motion slows down or accelerates and whether they expand with time, and possibly observe the generation of new features. Measurements with ZIMPOL, a rapid-switching imaging polarimeter on SPHERE, of scattering polarization can constrain the phase angle and thus the line-of-sight configuration of the features relative to the disk. The Atacama

Large Millimeter/submillimeter Array (ALMA) observations can improve constraints on the disk's residual gas content. Monitoring the flaring activity of AU Mic may allow us to test the link between the generation of features in the dust distribution and coronal mass ejections. Finally, H α differential imaging may reveal signs of accretion if there exist proto-planets in the system.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 23 January; accepted 28 August 2015.

- Smith, B. A. & Terrell, R. J. A circumstellar disk around β Pictoris. *Science* **226**, 1421–1424 (1984).
- Mouillet, D., Larwood, J. D., Papaloizou, J. C. B. & Lagrange, A. M. A planet on an inclined orbit as an explanation of the warp in the β Pictoris disc. *Mon. Not. R. Astron. Soc.* **292**, 896–904 (1997).
- Augereau, J. C., Nelson, R. P., Lagrange, A. M., Papaloizou, J. C. B. & Mouillet, D. Dynamical modeling of large scale asymmetries in the β Pictoris dust disk. *Astron. Astrophys.* **370**, 447–455 (2001).
- Lagrange, A.-M. *et al.* The position of β Pictoris b position relative to the debris disk. *Astron. Astrophys.* **542**, A40 (2012).
- Nesvold, E. R. & Kuchner, M. J. A SMACK model of colliding planetesimals and dust in the β Pictoris debris disk: thermal radiation and scattered light. Preprint at <http://arxiv.org/abs/1506.07187> (2015).
- Liu, M. C. Substructure in the circumstellar disk around the young star AU Microscopii. *Science* **305**, 1442–1444 (2004).
- Metchev, S. A., Eisner, J. A., Hillenbrand, L. A. & Wolf, S. Adaptive optics imaging of the AU Microscopii circumstellar disk: evidence for dynamical evolution. *Astrophys. J.* **622**, 451–462 (2005).
- Krist, J. E. *et al.* Hubble Space Telescope Advanced Camera for Surveys coronagraphic imaging of the AU Microscopii debris disk. *Astron. J.* **129**, 1008–1017 (2005).
- Fitzgerald, M. P., Kalas, P. G., Duchêne, G., Pinte, C. & Graham, J. R. The AU Microscopii debris disk: multiwavelength imaging and modeling. *Astrophys. J.* **670**, 536–556 (2007).
- Robinson, R. D., Linsky, J. L., Woodgate, B. E. & Timothy, J. G. Far-ultraviolet observations of flares on the dM0e star AU Microscopii. *Astrophys. J.* **554**, 368–382 (2001).
- Mamajek, E. E. & Bell, C. P. M. On the age of the β Pictoris moving group. *Mon. Not. R. Astron. Soc.* **445**, 2169–2180 (2014).
- Kalas, P., Liu, M. C. & Matthews, B. C. Discovery of a large dust disk around the nearby star AU Microscopii. *Science* **303**, 1990–1992 (2004).
- Augereau, J. C. & Beust, H. On the AU Microscopii debris disk. *Astron. Astrophys.* **455**, 987–999 (2006).
- Roberge, A., Weinberger, A. J., Redfield, S. & Feldman, P. D. Rapid dissipation of primordial gas from the AU Microscopii debris disk. *Astrophys. J.* **626**, L105–L108 (2005).
- Schneider, G. *et al.* Probing for exoplanets hiding in dusty debris disks: disk imaging, characterization, and exploration with HST/STIS multi-roll coronagraphy. *Astron. J.* **148**, 59 (2014).
- Beuzit, J.-L. *et al.* SPHERE: a 'Planet Finder' instrument for the VLT. In *Proc. SPIE Conf. Ser.*, **7014**, 18 (Society of Photo-Optical Instrumentation Engineers, 2008).

Acknowledgements SPHERE was built by a European consortium led by IPAG (France). SPHERE was funded by the ESO, with additional contributions from CNRS, MPA, INAF, FINES and NOVA. SPHERE also received funding from the European Commission FP6 and FP7 programmes as part of OPTICON under grant numbers RI3-Ct-2004-001566 (FP6), 226604 (FP7) and 312430 (FP7). French co-authors are supported by ANR-14-CE33-0018. Part of this work has been carried out within the framework of the National Centre for Competence in Research PlanetS supported by the Swiss National Science Foundation. C.T. and M.R.M. acknowledge the financial support of the SNSF. This study is based on observations from program 60.A-9249(C) at ESO Very Large Telescope and from program number 12228 made with the NASA/ESA Hubble Space Telescope, obtained at STScI, which is operated by AURA Inc. under NASA contract NAS 5-26555. We are also grateful to the ESO for releasing the commissioning data for publication. Finally, we thank P. Zarka, N. Meyer-Vernet, B. Stelzer and Q. Kral for discussions. J.S. is a NASA Postdoctoral Program Fellow.

Author Contributions A.B. reduced and analysed IRDIS data and wrote the paper. C.T. reduced the IRDIS and ZIMPOL data and contributed to the manuscript writing. A.B., C.T., A.-M.L., M.J., J.-C.A., G.S., C.G., J.D., D.M., T.H., C.D., C.G., J.O. and J.S. worked on the interpretation of the results. G.S., C.G., J.D., D.H., T.R. and J.W. re-reduced HST data. J.M., M.L., D. Mouillet, D. Mawet, J.H.G. and Z.W. operated the instrument at the telescope. A.-L.M. worked on the astrometric calibration. A.B., A.-M.L., M.L., D. Mouillet, T.H., J.-L.B., K.D., M.F., T.F., M.K., F.M., M.M., C.M., J.-F.S., H.M.S., M.T., S.U., F.V. and A.V. contributed to the instrument conception. All authors commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.B. (anthony.boccaletti@obspm.fr).

METHODS

Observations and data reduction. SPHERE is a highly specialized instrument dedicated to high-contrast imaging, built by a wide consortium of European laboratories and recently installed at the VLT¹⁶. It is based on the Sphere Adaptive optics for eXoplanet Observation (SAXO) extreme adaptive optics system, with a 41×41 actuator wavefront control. Several coronagraphic devices for stellar diffraction suppression are provided, including apodized Lyot coronagraphs.

AU Mic was observed on 11 August 2014, with the differential imaging camera IRDIS, in the J band for a total integration time of 2,560 s. IRDIS offers two square fields of view (about $12''$ each) on the same detector to allow for spectral differential imaging, but since the broadband J filter was used for both channels, they provided redundancy in this case. The star was masked with an apodized Lyot coronagraph of which the focal mask occults an area of 185 mas in diameter and the pupil mask transmits $\sim 67\%$ of the light.

Data reduction follows a standard procedure including cosmetics (correction for flat-field, bad pixels, dark current, and distortion). Individual frame registration is not required as the sequence is very stable and a dedicated hardware in SPHERE is taking care of the positioning of the star onto the coronagraph in real time, reaching 0.5 mas accuracy¹⁷.

We then took advantage of the field rotation during the observation ($\sim 77^\circ$) to suppress the residual starlight in coronagraphic images and reveal the faint scattered light from the debris disk via angular differential imaging (ADI). We explored several ADI techniques, including classical ADI, LOCI¹⁸ and KLIP¹⁹ with various parameter settings. The final images were obtained with a LOCI (Fig. 1e) frame selection criterion of 0.75 full-width at half-maximum (FWHM), and an optimization zone of 10,000 PSF footprints (using sectors of annuli 12 FWHM in the radial dimension), while the KLIP image (Fig. 1d) is calculated for separations shorter than 600 pixels ($7.35''$), and is built from the subtraction of 5 modes out of 160 (a conservative value to avoid strong attenuation of the disk). Since ADI techniques achieve their high contrast performance at the cost of flux losses to the disk image, which remain difficult to calibrate²⁰, we also reduced the data with less powerful but more conservative methods, such as frame-by-frame reference star subtraction or subtraction of an azimuthally averaged radial profile. Owing to the high quality and stability of the data, these methods performed similarly to the ADI techniques in terms of detecting the disk at separations larger than ~ 0.7 – $1.0''$. Doing so, the processed images reach a 5σ contrast as large as 4×10^{-6} at $\sim 0.5''$. All data reduction methods recover the disk features A–E consistently and at the same locations.

The limit of detection to point sources is presented in Extended Data Fig. 1, as measured within the disk using the method of fake point sources injection to calibrate for the self-subtraction inherent to ADI. The contrast achieved in the image at a projected separation of $1''$ would have enabled the detection of a planet with a mass 1 to 6 times that of Jupiter, depending which evolutionary and atmospheric model is considered^{21,22}, and assuming an age of 20 Myr. This threshold potentially lowers to a Saturn-mass object at a projected separation of $4''$ (about the location of the planetesimal belt), but the models are not reliable for such low mass.

The published images of the optical HST observations from 2010 and 2011 represent a combination of both epochs¹⁵, in which the strongest feature (B) was already identified. For the purpose of tracking our disk features A–E through time, these data were re-reduced to yield separate images for both epochs. Following the original recipe for data reduction (multi-roll PSF-template subtraction) augmented with an additional high-pass filtering (unsharp masking), we recover features B–E reliably in both epochs. In 2010, feature A resides inside the blind area resulting from the multi-roll technique. The HST images are obtained with STIS in a filter-less mode, the spectral range being set by the detector spectral response across a very broad band (200–1,100 nm).

For Figs 2a–c, the images shown in Fig. 1a–c were unsharp-masked on a spatial scale of $0.76''$. The main body of the disk was approximated as a brightness distribution with a broken linear horizontal profile (with a break at $3''$, the approximate radius of the source ring) and a Gaussian vertical profile. The linear trends were chosen on the basis of the disk's brightness profile along the midplane. This distribution was subtracted to reveal the inhomogeneous substructure. The same distribution was used for both HST epochs, preserving their extreme reliability. In Fig. 2d, a more aggressive asymmetric kernel of $0.76'' \times 0.25''$ was used for unsharp-masking on the two HST images to highlight sharp horizontal gradients that are suitable for visualizing the differential motion.

On 13 August 2014, we obtained a follow-up observation of AU Mic with ZIMPOL, the rapid-switching imaging polarimeter. A total of 1 h of integration time was taken in the I'-band filter (713–866 nm) in imaging mode (no polarimetry) with pupil tracking so as to allow for ADI data reduction. Since the high-sensitivity detector mode is currently only available in the slow polarimetry mode,

which does not support pupil tracking, the noisier high-gain detector mode was used. AU Mic was heavily saturated and produced some charge bleeding in the vertical direction, but the compromised region does not affect the disk detection. A total of 40° of field rotation was captured during the observation. After correcting for cosmetics as for the IRDIS data, we applied various ADI data reduction techniques to suppress the stellar halo. In Extended Data Fig. 2, we show the results for LOCI data reduction with 'conservative' parameter settings²³ including a frame selection criterion of $0.5 \times$ FWHM and an optimization area of 10,000 PSF footprints (same geometry as for IRDIS). The adaptive optics correction is more difficult at shorter wavelengths, and thus yields a lower Strehl ratio in the optical than in the infrared. On the other hand, the shorter wavelengths yield a higher angular resolution ($\lambda/D \approx 19$ mas in the I'-band as compared to about 32 mas in the J-band) and thus a greater potential to resolve fine structure. As Extended Data Fig. 2 demonstrates, the location and overall morphology of the A and B features as seen in the IRDIS images are very well reproduced in the ZIMPOL images, including the wave-like connection of feature A to the disk plane. Both images show an additional pattern in between feature B and the midplane. This structure may represent further wave-like features like A–E at a lower amplitude, but will require future investigation and modelling.

Disk morphology. From a morphological point of view, the southeast and northwest sides of the disk are very different. The former contains many structures above the midplane, while the latter is brighter, thinner and features an abrupt change of direction near $1.5''$. The disk PA, measured from north to east, is determined in both SPHERE and HST images using the method developed earlier for edge-on disks⁴. The image is rotated with an initial guess for the PA to place the disk midplane approximately horizontally and a profile function (Gaussian or Lorentzian) is fitted vertically to retrieve the midplane centroid versus the angular separation. We used regions where the disk contains as few features as possible ($3''$ to $6''$ here). The true disk PA is the image rotation for which the slope of the disk centroid is flat. The measurement is repeated separately for the two sides, since the AU Mic disk is highly asymmetric. We found $PA_{SE} = 129.5 \pm 0.2^\circ$ and $PA_{NW} = 311.2 \pm 0.3^\circ$ respectively for the southeast and northwest sides. Similarly in the HST images we obtained $PA_{SE} = 129.0 \pm 0.5^\circ$ and $PA_{NW} = 310.5 \pm 0.2^\circ$. Although the error is relatively large, the measurements from SPHERE and HST differ in the northwest side by $0.7 \pm 0.4^\circ$. We suspect that the determination of the PA in the southeast is in fact perturbed by the presence of the features appearing at different locations between 2010–2011 and 2014. Thus, we considered that the northwest side gives a more reliable measurement of PA so we compensated the HST image with a rotation of 0.7° to realign all the epochs. We note that for both HST and SPHERE the true north uncertainty is $\sim 0.1^\circ$, so the uncertainty on the disk PA is reflecting our ability to locate the disk midplane and is also possibly affected by the colour dependence of the grains. In addition, the two sides are clearly misaligned by $1.7 \pm 0.4^\circ$ in the SPHERE image and by $1.5 \pm 0.5^\circ$ in the HST image. Once the disk PA is set, the centroid of the disk cross-section versus separation defines the disk spine in which the features are visible as excursions from the midplane (Extended Data Fig. 3). This spine includes both the main disk and the features, which explains that they may appear in Extended Data Fig. 3 at different elevations than in Fig. 1. To register the radial locations of features we used a model profile combining a Gaussian and a first-order polynomial in some delimited regions (red lines in Extended Data Fig. 3). The measurement is repeated for various data reductions, including PA uncertainty, to estimate the errors on the location of the features. The registration of features and associated errors are listed in Extended Data Table 1.

Finally, we also found that the disk spine shows an excursion of $0.07''$ (equivalent to ~ 5 – 6 pixels) southwest to the star inside a radius of 0.6 – $0.7''$, a characteristic that is clearly seen in a zoomed image (Extended Data Fig. 4). It is unlikely to be a result of ADI bias, which is expected to be symmetrical about the disk midplane. Similar excursions were observed in a number of debris disks; these excursions could represent the opening of each disk's source ring as viewed at an inclination close to, but not equal to, 90° . In such a situation, anisotropic forward-scattering is expected to render the near-side edge of the ring much brighter than the far-side edge, which accounts for the asymmetry. A complete analysis of the disk photometry is deferred to future work, since a careful modelling of ADI bias effects is crucial in that case, especially close to the star²⁰.

With the most aggressive algorithms (those that remove the starlight most efficiently, like KLIP and LOCI) the three features closest to the star appear as arches; that is, the structures are clearly separated from the midplane by a void of scattered light. As a qualitative 'sanity check', fake bumps were added to the data inside the disk midplane to investigate qualitatively whether ADI could produce a depletion between the midplane and the top of the bump, mimicking arches. We found no such effects and therefore conclude that this is probably a real characteristic, to be confirmed with deeper follow-up observations.

As a complement to Figs 3 and 4, we have plotted the stellocentric distance versus time for each feature in Extended Data Fig. 5. The structures are well aligned

over the three epochs, error bars being smaller than the plotted symbols in some cases. Once the data points are fitted with linear trends and extended back in time, three out of five features (A, B, C) lie on nearly parallel tracks, and suggest a timeframe of ~ 15 years (where lines intersect the y axis). In fact, the observed structures are necessarily recent, otherwise they would have propagated and smeared all around the star as a result of secular evolution. Brightness asymmetries reported in the literature in 2004 may coincide with the tracks for features C and D, though it is difficult to determine reliably whether they are the same features since they are seen as intensity variations rather than excursions from the midplane.

Physical interpretation of disk features. A majority of known debris disks exhibit structural features such as eccentricities, warps and brightness asymmetries, which are assumed to be induced by planets via secular gravitational perturbation. However, such features either appear static over observational timescales or remain coupled to the Keplerian motion of the disk, which is incompatible with the fast motion measured for two or three of the five features observed. There are mandatory observational facts with which a physical interpretation must comply, at least qualitatively, which are: (1) spatial localization of the features on one side of the disk and above the midplane, (2) timeframe for the evolution, (3) increase of projected speeds at larger projected separations, (4) larger projected radial widths away from the star, (5) increase of intensities at shorter projected separations, and (6) variable elevations.

Although a number of mechanisms occur in massive protoplanetary disks that can affect the dust distribution and generate structures with speeds of a few to a few tens of kilometres per second, they rely on the presence of gas. Although some debris disks retain a large amount of gas²⁴, such gas is probably a low-mass component in the AU Mic system¹⁴ compared to the estimated total mass of dust⁹. For these reasons gas-induced scenarios (such as radiation-driven disk wind and protostellar jets) are considered unlikely here.

One possible assumption would be that the measured speeds represent the phase speed of a pattern propagating through the disk, which could greatly exceed the physical speed of the constituent disk particles. Indeed, protoplanetary disks may exhibit spiral density waves whose outer arms ‘travel’ at super-Keplerian speeds, as a response to gravitational instabilities or planets orbiting inside the disk²⁵. Given AU Mic’s youth, it must have dispersed its primordial gas only recently; thus, some disk structures could conceivably have survived as ‘fossils’. Whether this is physically plausible remains to be investigated. Resonances can induce wave-like structure even in gas-less disks. Saturn’s rings feature edge waves along the orbits of embedded moons, although they follow Keplerian orbits²⁶. Lindblad resonances, on the other hand, produce spirals phase-locked to a planet, which exhibit super-Keplerian phase speeds. However, a spiral would have to ‘wrap’ around the star several times to reproduce the observed train of features on the southeastern side, which is at odds with the lack of features on the north-western side.

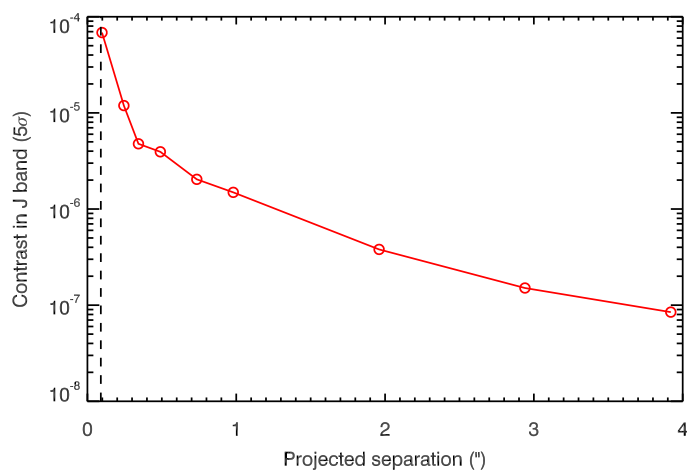
Local intensity enhancements on one side only could be interpreted as a series of concentric eccentric rings resulting from massive collisions of asteroid-like objects²⁷. However, the typical timescale to produce several eccentric rings is of the order of 100 years, too long compared to our measurements for the moving structures in the AU Mic disk.

Rather than phase speed, the observed motion may represent physical motion at super-Keplerian velocities. Dust blowout by stellar radiation or wind constitutes an integral part of the mechanism that produces debris disks, and is well capable of boosting small grains to escape speeds. Given AU Mic’s high activity level, flares

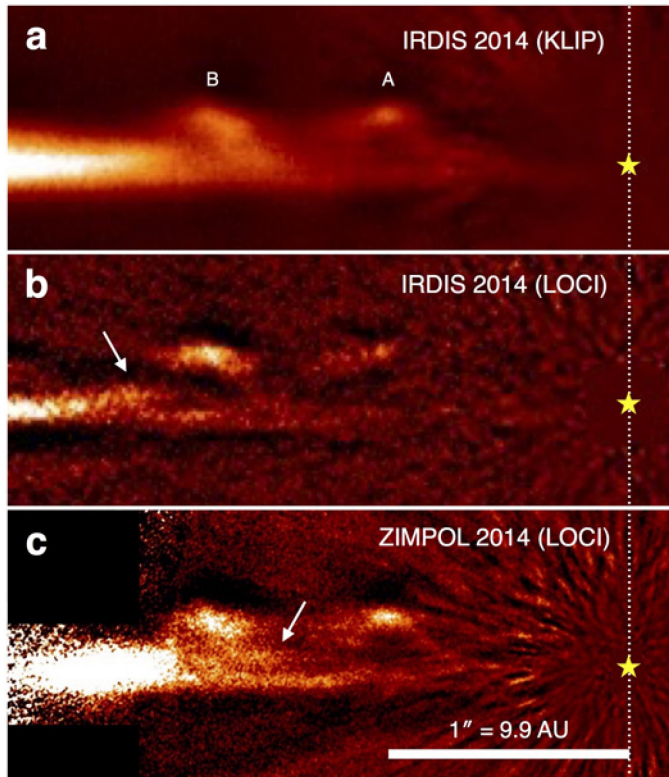
from coronal mass ejections could occasionally impact the planetesimal ring and produce distinct dust clouds at different azimuths. A warped ring of planetesimals as in β Pictoris⁵ could account for the elevation. Owing to anisotropic scattering, the near-side clouds could appear bright while those on the far side remain undetected, explaining the one-sided apparent distribution. Similarly, the interaction of episodic flares with a planet’s magnetosphere or a dusty circumplanetary disk, on a Keplerian orbit, may explain the spatial localization of the features as a train of dust cloud²⁸. Circumplanetary disks are also capable of releasing outflows²⁹. In both of these scenarios, the combination of orbital motion of the dust source and the outward force would explain the velocity dispersion shown in Fig. 4.

In the planetary outflow scenario, given that features A and E could have been released approximately 15 years apart (Extended Data Fig. 5) and that projected speeds vary from $\sim 4 \text{ km s}^{-1}$ to 10 km s^{-1} , we can constrain the minimal separation of a planet to $\sim 10\text{--}15 \text{ AU}$. On the other hand, dust clearing observed at distances closer than $\sim 35\text{--}40 \text{ AU}$ could be the result of a planet orbiting inside the planetesimal belt. Therefore, in the range $10\text{--}40 \text{ AU}$, where a hypothetical planet may reside, the SPHERE data reach a contrast of 1×10^{-6} to 8×10^{-8} , which, for the DUSTY model²¹, provides an upper limit at 6 and 3.5 Jupiter masses, respectively. **Code availability.** Data reductions are performed with IDL and custom routines (including IDP3 available from the Mikulski Archive for Space Telescopes at STScI).

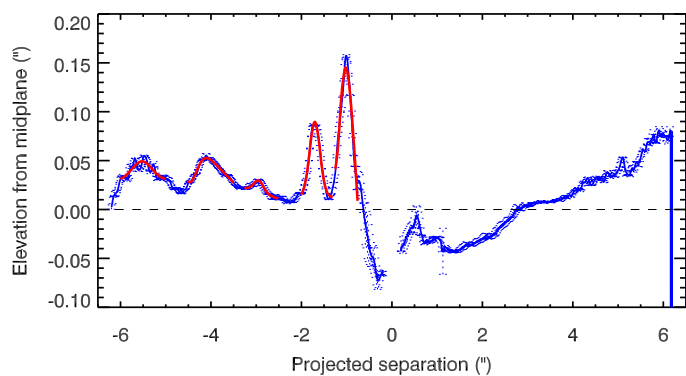
17. Baudoz, P. *et al.* The differential tip-tilt sensor of SPHERE. In *Proc. SPIE Conf. Ser.* **7735**, 5 (Society of Photo-Optical Instrumentation Engineers, 2010).
18. Lafrenière, D., Marois, C., Doyon, R., Nadeau, D. & Artigau, E. A new algorithm for point-spread function subtraction in high-contrast imaging: a demonstration with angular differential imaging. *Astrophys. J.* **660**, 770–780 (2007).
19. Soummer, R., Pueyo, L. & Larkin, J. Detection and characterization of exoplanets and disks using projections on Karhunen–Loève eigenimages. *Astrophys. J.* **755**, L28 (2012).
20. Milli, J. *et al.* Impact of angular differential imaging on circumstellar disk images. *Astron. Astrophys.* **545**, A111 (2012).
21. Chabrier, G., Baraffe, I., Allard, F. & Hauschildt, P. Evolutionary models for very low-mass stars and brown dwarfs with dusty atmospheres. *Astrophys. J.* **542**, 464–472 (2000).
22. Baraffe, I., Chabrier, G., Barman, T. S., Allard, F. & Hauschildt, P. H. Evolutionary models for cool brown dwarfs and extrasolar giant planets. The case of HD 209458. *Astron. Astrophys.* **402**, 701–712 (2003).
23. Thalmann, C. *et al.* Imaging of a transitional disk gap in reflected light: indications of planet formation around the young solar analog LkCa 15. *Astrophys. J.* **718**, L87–L91 (2010).
24. Moór, A. *et al.* ALMA continuum observations of a 30 Myr old gaseous debris disk around HD 21997. *Astrophys. J.* **777**, L25 (2013).
25. Muto, T. The structure of a self-gravitating protoplanetary disk and its implications for direct imaging observations. *Astrophys. J.* **739**, 10 (2011).
26. Weiss, J. W., Porco, C. C. & Tiscareno, M. S. Ring edge waves and the masses of nearby satellites. *Astron. J.* **138**, 272–286 (2009).
27. Kral, Q., Thébaud, P., Augereau, J. C., Boccaletti, A. & Charnoz, S. Signatures of massive collisions in debris discs. A self-consistent numerical model. *Astron. Astrophys.* **573**, A39 (2015).
28. Kivelson, M. G. & Southwood, D. J. Dynamical consequences of two modes of centrifugal instability in Jupiter’s outer magnetosphere. *J. Geophys. Res.* **110**, A12209 (2005).
29. Fendt, C. Magnetically driven outflows from Jovian circum-planetary accretion disks. *Astron. Astrophys.* **411**, 623–635 (2003).



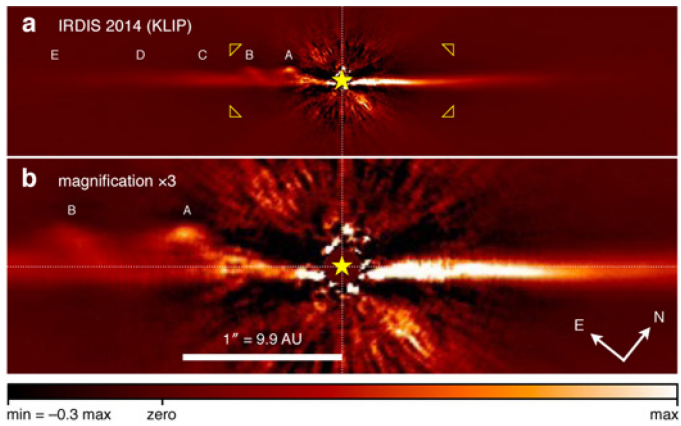
Extended Data Figure 1 | Limit of detection to point sources. The contrast is measured at 5σ using fake planets introduced to the data at discrete positions (circles) along the disk midplane to account for the self-subtraction of the ADI/KLIP algorithm. The dashed line defines the edge of the coronagraphic mask at $0.09''$.



Extended Data Figure 2 | Comparison of IRDIS and ZIMPOL images. **a** and **b** show zoomed-in regions of the KLIP and LOCI reductions of the IRDIS infrared data, whereas **c** is taken from the conservative LOCI reduction of the ZIMPOL optical data. Features A and B are reproduced accurately in the ZIMPOL data. An additional substructure between feature B and the midplane is also detected, as indicated by arrows. The yellow star symbol indicates the position of the star.

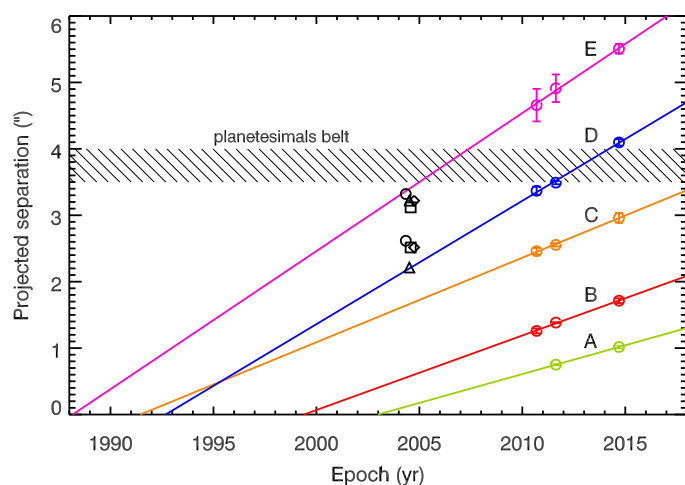


Extended Data Figure 3 | Spine of the disk measured in SPHERE IRDIS data. The spine is measured using several reductions (noADI, ADI, KLIP) of the SPHERE IRDIS 2014 data. Average values and dispersions (error bars) are plotted as a blue line. For each region where a local maxima is identified, a Gaussian + first-order polynomial model is fitted in order to register precisely the five features.



Extended Data Figure 4 | Central part of the SPHERE IRDIS image.

a shows a $12''$ field of view of the SPHERE IRDIS image processed with the KLIP algorithm and **b** is a magnified version to indicate the bow-like deviation of the disk to the southeast in the central area (for separations shorter than $\sim 0.7''$). The horizontal dotted lines indicate the disk midplane.



Extended Data Figure 5 | Positions of the disk features over time. The positions of the features measured in the SPHERE and HST images are plotted as circles together with peak-to-valley error bars (in some cases, the errors are smaller than the symbol size). Linear fits on these three epochs illustrate the possible track of each feature. The black symbols show the location at which various inhomogeneities were reported in the literature, on the basis of older data^{6–9}. The colour coding is the same as in Fig. 4.

Extended Data Table 1 | Registration of features

| Epoch | feature A | feature B | feature C | feature D | feature E |
|-------------|---------------------|---------------------|---------------------|---------------------|---------------------|
| 2010 HST | – | $1.259 \pm 0.037''$ | $2.459 \pm 0.049''$ | $3.369 \pm 0.061''$ | $4.658 \pm 0.245''$ |
| 2011 HST | $0.750 \pm 0.012''$ | $1.384 \pm 0.012''$ | $2.554 \pm 0.025''$ | $3.491 \pm 0.025''$ | $4.912 \pm 0.208''$ |
| 2014 SPHERE | $1.017 \pm 0.025''$ | $1.714 \pm 0.037''$ | $2.961 \pm 0.073''$ | $4.096 \pm 0.049''$ | $5.508 \pm 0.074''$ |

Two-channel Kondo effect and renormalization flow with macroscopic quantum charge states

Z. Iftikhar¹, S. Jezouin¹, A. Anthore^{1,2}, U. Gennser¹, F. D. Parmentier¹, A. Cavanna¹ & F. Pierre¹

Many-body correlations and macroscopic quantum behaviours are fascinating condensed matter problems. A powerful test-bed for the many-body concepts and methods is the Kondo effect^{1,2}, which entails the coupling of a quantum impurity to a continuum of states. It is central in highly correlated systems^{3–5} and can be explored with tunable nanostructures^{6–9}. Although Kondo physics is usually associated with the hybridization of itinerant electrons with microscopic magnetic moments¹⁰, theory predicts that it can arise whenever degenerate quantum states are coupled to a continuum^{4,11–14}. Here we demonstrate the previously elusive ‘charge’ Kondo effect in a hybrid metal–semiconductor implementation of a single-electron transistor, with a quantum pseudospin of 1/2 constituted by two degenerate macroscopic charge states of a metallic island^{11,15–20}. In contrast to other Kondo nanostructures, each conduction channel connecting the island to an electrode constitutes a distinct and fully tunable Kondo channel¹¹, thereby providing unprecedented access to the two-channel Kondo effect and a clear path to multi-channel Kondo physics^{1,4,21,22}. Using a weakly coupled probe, we find the renormalization flow, as temperature is reduced, of two Kondo channels competing to screen the charge pseudospin. This provides a direct view of how the predicted quantum phase transition develops across the symmetric quantum critical point^{4,21}. Detuning the pseudospin away from degeneracy, we demonstrate, on a fully characterized device, quantitative agreement with the predictions for the finite-temperature crossover from quantum criticality¹⁷.

In previous experimental investigations, the Kondo quantum impurity was of a microscopic nature and mostly associated with spin^{6,7,9,23–25}, orbital^{8,26} or possibly structural degrees of freedom^{4,27}. In the ‘charge’ Kondo effect^{11,16,17}, it is a pseudospin of 1/2 consisting of two degenerate states of a macroscopic quantum variable, namely, the electrical charge of a metallic island comprising several billion electrons. The role of the electrons’ spin (\uparrow or \downarrow) in the original spin Kondo problem¹⁰ is played by the electrons’ location, which can be in the island (\uparrow) or elsewhere (\downarrow). Accordingly, the charge pseudospin flips when electrons are transferred into and out of the island. The Kondo channels, each coupling the Kondo impurity (pseudospin) with a distinct electron continuum, directly equate with the different electrical conduction channels connected to the island (distinguishing between those associated with different values of the real electron spin). In contrast, the electrical channels in previous Kondo nanostructures normally merged into a single Kondo channel (except in the ingenious implementation of ref. 9), owing to cooperative spin-flip processes involving charge transfers between continuums. Furthermore, the charge pseudospin energy splitting, adjusted by detuning the island from degeneracy with a gate voltage, is fully equivalent to the Zeeman splitting of a magnetic Kondo impurity. Finally, of practical importance, the macroscopic charge pseudospin allows for large channel distances, and thereby enables full and independent control as well as the *in situ* characterization of every Kondo parameter, giving access to direct comparisons with theory.

Here we investigate a nanostructure designed to display the two-channel ‘charge’ Kondo effect^{11,16,17}. The device (Fig. 1a) is a hybrid metal–semiconductor single-electron transistor (SET) with additional characterization probes. It essentially consists of a central metallic island (shown bright in Fig. 1a), with a continuous electronic density

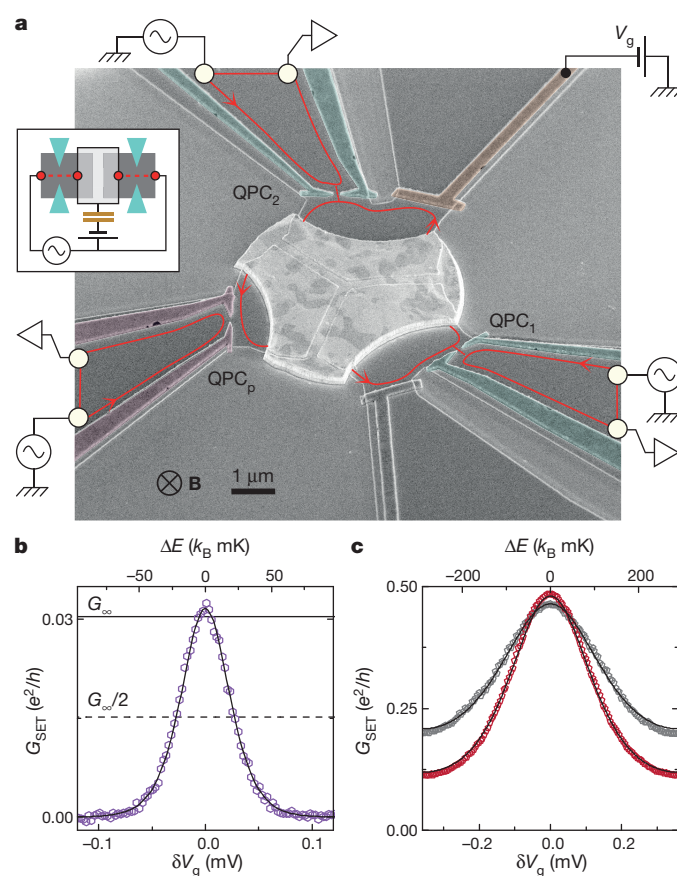


Figure 1 | Hybrid metal–semiconductor single-electron transistor.

a, Coloured picture of the sample (schematic in inset), consisting of a central metallic island (shown bright) connected to large electrodes (white circles) through the quantum point contacts QPC_{1,2} formed in a buried 2DEG (darker grey). The lateral continuous gates and QPC_p are used, respectively, to characterize the ‘intrinsic’ and ‘*in situ*’ (renormalized) conductances of QPC_{1,2}. The magnetic field $B \approx 3.9$ T corresponds to the integer quantum Hall regime, with the current propagating along spin-polarized edge channels (red lines) in the direction indicated by arrows. **b**, **c**, Kondo renormalized Coulomb peaks. Shown are measured SET conductance (symbols) versus gate voltage V_g (pseudospin energy splitting ΔE), for symmetric QPC_{1,2} set to $\tau_{1,2} \approx 0.06$ (**b**) at $T \approx 11.5$ mK or $\tau_{1,2} \approx 0.93$ (**c**) at $T \approx 11.5$ mK (red) and 22 mK (grey). Continuous lines are theoretical predictions (see main text, Methods). The agreement between data and theory in **c** establishes the predictions for the crossover from quantum critical behaviour as a function of ΔE (main text).

¹CNRS, Laboratoire de Photonique et de Nanostructures (LPN), 91460 Marcoussis, France. ²Univ Paris Diderot, Sorbonne Paris Cité, LPN, 91460 Marcoussis, France.

of states, connected to large electrodes through two quantum point contacts (QPC_{1,2}), each tuned to a single conduction channel. The QPCs are formed in a Ga(Al)As two-dimensional electron gas (2DEG) by the field effect using split gates. The 2DEG is further confined by etching (to the darker grey areas in Fig. 1a), and electrically connected to the metallic island by thermal annealing. The lateral continuous gates are used to extract the ‘intrinsic’ transmission probabilities $\tau_{1,2}$ characterizing QPC_{1,2}, respectively, by short-circuiting the central island. The capacitively coupled gate voltage V_g controls the energy difference between the island charge states. When set to weak coupling, QPC_p gives us access separately to the ‘*in situ*’ conductances $G_{1,2}$ of QPC_{1,2}, respectively. Except when specifically indicated, QPC_p is disconnected.

The experiment is performed down to an electronic temperature $T \approx 11.5$ mK (Methods), in a perpendicular magnetic field $B \approx 3.9$ T that breaks the spin degeneracy and corresponds to the integer quantum Hall effect at filling factor 2. In this regime, the current flows along two (spin-polarized) chiral edge channels. Red lines in Fig. 1a represent the outer channel, closest to the edge, with the propagation direction indicated by arrows. It is partially transmitted across QPC_{1,2}, whereas the inner channel (not shown) is fully reflected and can be ignored.

We now review the main requirements for mapping the physics of this device to the two-channel Kondo (2CK) problem. First, the typical electronic level spacing δ in the metallic island should be much smaller than the thermal energy: $\delta \ll k_B T$, with k_B the Boltzmann constant^{16,17}. We estimate $\delta \approx k_B \times 0.2$ μ K (Methods), four orders of magnitude smaller than $k_B T$. Second, the charging energy $E_C = e^2/2C$, with e the electron charge and C the overall island capacitance, should be larger than $k_B T$ to reduce the accessible charge states to a pseudospin of 1/2. We obtain from standard Coulomb diamond analysis $E_C \approx k_B \times 290$ mK (Methods). Third, the metallic island should be in nearly perfect contact with the 2DEG, in particular to avoid resonances involving the 2DEG–metal interface. We find that the outer edge channel is fully transmitted into the metallic island, with a reflection probability smaller than 0.05% (Methods). Finally, QPC_{1,2} should implement point-like contacts, with a small energy dependence of $\tau_{1,2}$. For the experimental set points, we find using the lateral characterization gates that $\tau_{1,2}$ increase monotonically with energy by at most 11% up to $2E_C$ (Methods). Together, the last two tests rule out any resonant effects.

From the influence of the charge states’ energy splitting on conductance, we observe first indications of 2CK effects and establish that the measurements are performed in a regime where this physics is expected. The measured conductance G_{SET} of the QPC₁–island–QPC₂ SET is plotted as symbols versus gate voltage for symmetric QPCs set in the tunnel and weak-backscattering regimes, $\tau \equiv \tau_1 \approx \tau_2 \approx 0.06$ ($T \approx 11.5$ mK) and 0.93 ($T \approx 11.5$ and 22 mK), in Fig. 1b, c, respectively. The conductance exhibits periodic peaks located at successive charge degeneracy points (one full period $\Delta \approx 0.72$ mV is shown in Fig. 1c, Methods).

The tunnel data in Fig. 1b are compared with the prediction for incoherent sequential tunnelling events²⁸

$$G_{\text{SET}} = \frac{G_\infty}{2} \frac{2E_C(\delta V_g/A)/k_B T}{\sinh(2E_C(\delta V_g/A)/k_B T)} \quad (1)$$

with $G_\infty = (e^2/h)/(\tau_1^{-1} + \tau_2^{-1})$ the ‘classical’ SET conductance and h the Planck constant. We find that the data (symbols) can be accurately reproduced with a fit temperature of 10 mK (continuous line in Fig. 1b), slightly smaller than but compatible with $T \approx 11.5 \pm 1.5$ mK. However, the maximum peak conductance is much higher than the standard prediction $G_\infty/2$ (dashed line), and G_∞ was left as a free fit parameter. Such an increase is expected from the Kondo renormalization of the conductance, even for relatively low characteristic Kondo temperature scales $T_K \ll T$. In this limit, equation (1) is predicted to provide a good approximation when substituting G_∞ by $\sim \log^{-2}(T/\alpha T_K)$, with α a

numerical factor¹⁷ (Methods). Assuming $\tau \ll 1$, the Kondo temperature reads¹⁷

$$T_K^{\tau \ll 1} \approx (E_C/k_B) \exp(-\pi^2/\sqrt{4\tau}) \quad (2)$$

In the opposite limit of weak-backscattering, the 2CK physics is expected to be well developed. We find that the $\tau \approx 0.93$ data (Fig. 1c, symbols) are accurately reproduced, quantitatively and without fit parameters, by the predictions (lines) from the theoretical framework where the Kondo mapping is established¹⁷ (Methods).

With these indications of 2CK effects, we now provide direct experimental evidence of Kondo physics from the temperature dependence $G_{\text{SET}}(T)$ at the charge degeneracy point, with QPC_{1,2} remaining symmetric.

In standard metallic SETs, with many opaque conduction channels, the peak conductance monotonically decreases from its high-temperature classical value G_∞ as the temperature is reduced²⁹. In stark contrast, we find that $G_{\text{SET}}(\delta V_g = 0)$ increases as the temperature is reduced and, at $T \approx 11.5$ mK, always exceeds the classical conductance G_∞ by up to nearly 30% (Fig. 2a). Note that the separately characterized intrinsic energy dependencies of $\tau_{1,2}$ correspond to an opposite decrease of G_{SET} smaller than 1% for $T \lesssim 80$ mK. Remarkably, the conductance increase is logarithmic in T (continuous lines, for $T \lesssim 80$ mK), which is a typical signature of the Kondo effect.

A characteristic of Kondo systems, arising from renormalization group physics², is that they follow universal scaling laws. We demonstrate that the conductance data at $T \lesssim 80$ mK can be rescaled into a single curve $G_{\text{SET}}(T, \tau) = G_{\text{SET}}(T/T_K)$, and that the extracted $T_K(\tau)$ agrees with the theoretical prediction for the Kondo temperature (Fig. 2b). The simple rescaling procedure (Methods) relies on G_{SET} overlapping for different τ , and on the prediction $G_{\text{SET}}(T/T_K \gg 1) \propto \log^{-2}(T/\alpha T_K)$ (violet dashed line in Fig. 2b). The $T \ll T_K$ prediction $e^2/2h - G_{\text{SET}} \propto T/T_K$ is displayed as a red short-dashed line¹⁷ in Fig. 2b (Methods). The experimental scaling law covers an unprecedented range of T/T_K and most of $G_{\text{SET}} \in [0, 0.5]e^2/h$, thanks to the fully and independently tunable τ . Given the important G_{SET} overlaps, involving up to three successive values of τ , the rescaling accuracy provides a stringent test of the universal scaling law hypothesis. This conclusion is further established by comparing extracted $T_K(\tau)$ (symbols in inset in Fig. 2b) with the predictions derived at $1 - \tau \ll 1$ (Methods, red short-dashed line in inset in Fig. 2b), $\tau \ll 1$ (equation (2), violet dashed line in inset in Fig. 2b) and its generalization tested numerically (continuous line in inset in Fig. 2b)²⁰

$$T_K^{\text{num}} \approx (E_C/k_B) t \rho \exp(-\pi/(4t\rho)) \quad (3)$$

where $\pi t \rho \approx \sqrt{2(1 - \sqrt{1 - \tau})/\tau} - 1$. Adjusting the unknown theoretical prefactor to match $T_K(\tau \ll 1)$ or $T_K(1 - \tau \ll 1)$, we find an overall agreement over the whole range $\tau \in [0, 1]$.

With the ‘charge’ Kondo effect established, we turn to exploring the 2CK physics, which originates from the channels’ competition to screen the (pseudo)spin-1/2. Two symmetric Kondo channels are expected to flow, as $T \rightarrow 0$, towards the so-called¹⁷ strong-coupling fixed point characterized in the ‘charge’ Kondo implementation by two ballistic conduction channels ($G_{1,2} \rightarrow e^2/h$). In contrast to the one-channel Kondo (1CK) effect, this produces an over-screening of the pseudospin and, consequently, a non-Fermi liquid 2CK state with collective low-energy excitations²¹. The 2CK state is predicted to be unstable with an energy splitting of the pseudospin and with channel asymmetry, resulting in a $T = 0$ quantum phase transition²¹. Indeed, in the presence of an asymmetry, the most strongly coupled Kondo channel takes over, fully screening the pseudospin-1/2 at low temperatures and thereby hiding (decoupling) it from the other channel (see ref. 9 for first evidence of such a decoupling with a specific spin Kondo nanostructure³⁰). From the quantum phase transition perspective, the 2CK non-Fermi liquid character appears as a general consequence of the divergent correlations near the quantum critical point (symmetric and at degeneracy)¹. At finite $T \lesssim T_K$, the quantum critical (non-Fermi liquid) behaviour is preserved for a range of channel asymmetries and pseudospin energy

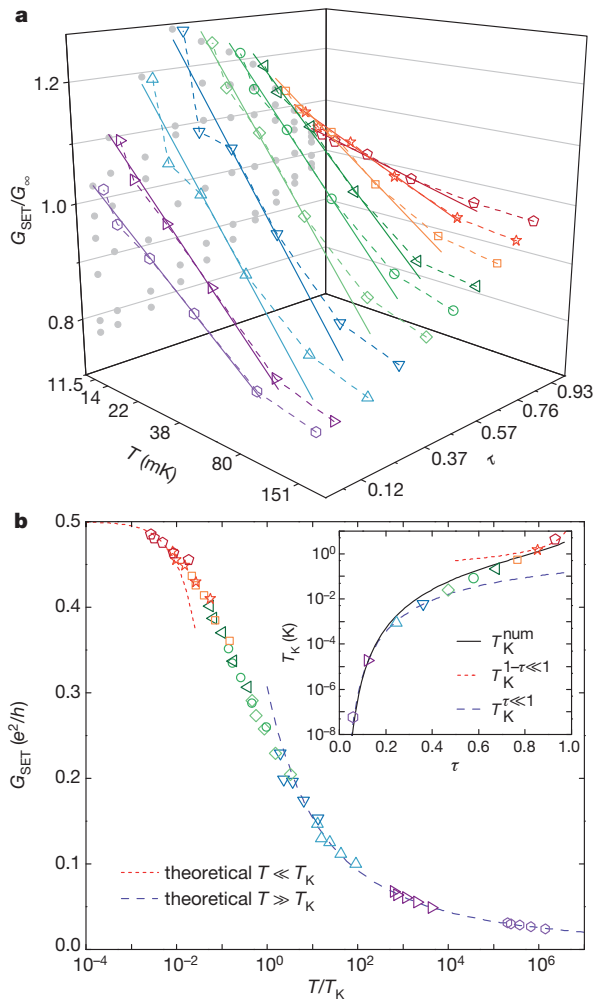


Figure 2 | Observation of the ‘charge’ Kondo effect. **a**, The normalized SET conductance $G_{\text{SET}}/G_{\infty}$, at charge degeneracy ($\delta V_g = 0$) and for symmetric QPC_{1,2}, is plotted as symbols versus the temperature on a logarithmic scale for different values of $\tau \equiv \tau_1 \approx \tau_2$. Continuous straight lines are guides to the eye proportional to $\log(T)$. The grey dots are the orthogonal projections of the different temperature measurements onto the plane $(\tau, G_{\text{SET}}/G_{\infty})$. **b**, The data in **a** at $T \leq 80$ mK, rescaled in temperature into a universal conductance curve (symbols). The violet dashed line displays the theoretical $T \gg T_K$ prediction $G_{\text{SET}} \propto \log^{-2}(T/\alpha T_K)$. The red short-dashed line displays the $T \ll T_K$ prediction $e^2/2h - G_{\text{SET}} \propto T/T_K$ (Methods). Inset, the extracted scaling parameter $T_K(\tau)$ (symbols) is compared to theoretical predictions (see equations (2) and (3) for the definitions of the Kondo temperatures $T_K^{\tau \ll 1}$ and $T_K^{\tau \gg 1}$, and Methods for $T_K^{1-\tau \ll 1}$).

splittings, which narrows down as T is reduced. Consequently, a non-Fermi to Fermi liquid crossover takes place^{4,21,22}.

The data from symmetric QPC configurations (Figs 1b, c, 2b) already reveal information on the 2CK physics. First, the experimental scaling law (Fig. 2b) shows that two symmetric Kondo channels flow monotonically towards the expected strong-coupling fixed point ($2G_{\text{SET}} \approx G_1 \approx G_2 \rightarrow e^2/h$ as $T \rightarrow 0$). Note that for $N \geq 3$ Kondo channels, the predicted symmetric fixed point is different^{12,17,21}. Second, the crossover from quantum critical to Fermi liquid behaviour with the pseudospin energy splitting $\Delta E = 2E_C \delta V_g/A$ is explored in Fig. 1c. Starting from a well-developed 2CK state at $\delta V_g = 0$ ($T/T_K \approx 0.003$ and 0.005), the SET conductance progressively moves away from the strong coupling fixed point $e^2/2h$ and, at sufficiently large ΔE , decreases as the temperature is reduced from 22 to 11.5 mK. Remarkably, the demonstrated agreement between data and theory for arbitrary δV_g validates quantitatively the theoretical description¹⁷ of the crossover. In particular, the crossover energy scale

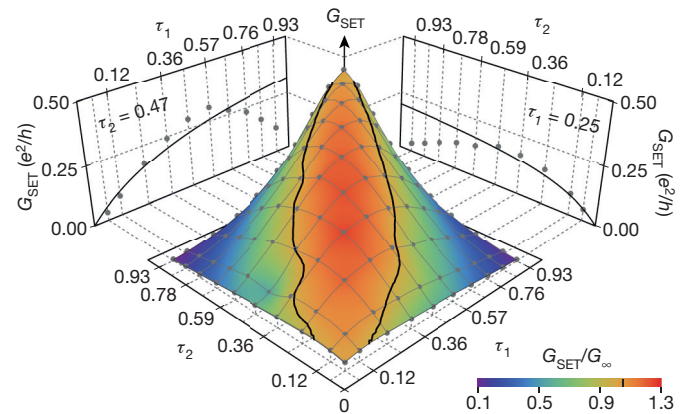


Figure 3 | Interplay of two Kondo channels revealed by tuning the asymmetry. Main plot, the SET conductance at charge degeneracy and $T = 11.5$ mK is displayed (symbols) versus QPC_{1,2} ‘intrinsic’ transmission probabilities, $\tau_{1,2}$. Narrow grey lines connect data points with the setting of one QPC fixed while the other is changed. The colour code represents $G_{\text{SET}}/G_{\infty}$ (black lines indicate $G_{\text{SET}} = G_{\infty}$). Lateral panels represent the same data (symbols) for a fixed value of $\tau_2 \approx 0.47$ (left) or $\tau_1 \approx 0.25$ (right), together with G_{∞} (continuous line).

(such that $G_{\text{SET}} = 0.5e^2/2h$) increases with T closely following the generic expectation $k_B T_K \sqrt{T/T_K}$ (Methods; we are preparing a thorough study of the crossover).

We provide evidence of the channels’ competition by exploring the effect of QPC asymmetry on G_{SET} . Symbols in Fig. 3 represent $G_{\text{SET}}(\tau_1, \tau_2)$ measured at $T \approx 11.5$ mK at the charge degeneracy point, while the colour code corresponds to the ratio $G_{\text{SET}}/G_{\infty}$. Note that it is only for nearly symmetric QPC_{1,2} that G_{SET} exceeds the ‘classical’ value G_{∞} (black continuous lines in Fig. 3). The stronger G_{SET} renormalization for symmetric QPCs indicates that they influence each other. Strikingly, G_{SET} exhibits a maximum and then decreases as the transmission probability of one QPC is continuously increased with the other fixed (symbols in lateral panels in Fig. 3). This non-monotonic behaviour demonstrates that the two QPCs are not independently renormalized, and validates expectations for two competing Kondo channels in series^{17,22}.

The 2CK phenomenology is directly revealed by the Kondo renormalization flow of the channels’ coupling, as temperature is reduced. It is experimentally characterized by the (renormalized) ‘*in situ*’ conductances $G_{1,2}$. We extract G_1 and G_2 separately by slightly opening QPC_p, with $G_p \ll G_{1,2}$ in order to minimize its effect (Methods). Figure 4 displays the (G_1, G_2) renormalization flow for $T \approx \{80, 38, 22, 14\}$ mK. The continuous lines connect data points (symbols) obtained for identical (τ_1, τ_2) , with an arrow indicating the flow direction and a colour corresponding to $|\tau_1 - \tau_2|$ ($\tau_{1,2} \lesssim 0.12$ and $|\tau_1 - \tau_2| \gtrsim 0.57$ are not included, owing to the small signal to noise).

First, note that the flow of asymmetric $G_{1,2}$ away from the symmetric line plainly exposes the development of the predicted quantum phase transition across the symmetric quantum critical point^{4,17,21,22}. Second, the renormalization flow also displays the predicted crossover from 2CK to 1CK behaviour. The 2CK zone of influence, shown as a grey background in Fig. 4, is characterized by an increase of both G_1 and G_2 as T is reduced. This occurs for $G_{1,2} \lesssim 0.5e^2/h$ (that is, $T \lesssim T_K$) or for relatively symmetric $G_{1,2}$. The 1CK zone of influence, shown as a light grey background in Fig. 4, is characterized by the reduction of the smallest ‘*in situ*’ conductance as T is lowered, while the largest further increases until reaching $\sim e^2/h$. This occurs for asymmetric $G_{1,2}$ and only if the largest ‘*in situ*’ conductance is above $\sim 0.5e^2/h$, corresponding to an important screening of the pseudospin. Note that the limit of one perfectly ballistic QPC was previously investigated in the context of dynamical Coulomb blockade³¹.

Further information is disclosed by the experimental renormalization flow, including the temperature evolution of channel asymmetry.

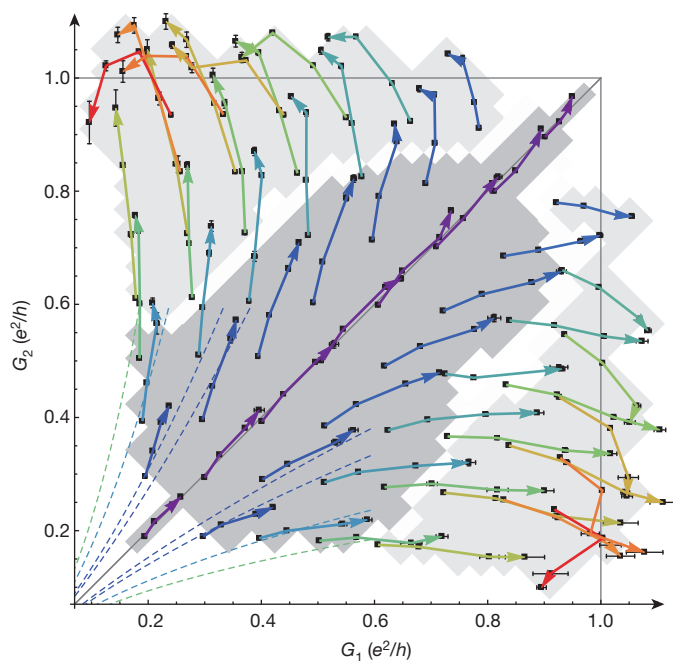


Figure 4 | Two-channel renormalization flow. The ‘*in situ*’ (Kondo renormalized) conductances (G_1, G_2) measured at $T \approx \{80, 38, 22, 14\}$ mK at charge degeneracy are displayed as symbols, with a line connecting different temperatures of the same QPC_{1,2} setting (characterized by the ‘intrinsic’, un-renormalized, $\tau_{1,2}$), and a colour code associated with $|\tau_1 - \tau_2|$ (from purple for $|\tau_1 - \tau_2| = 0$, to red for $|\tau_1 - \tau_2| = 0.57$). The arrows pointing to the 14 mK conductance data points show the flow direction for decreasing temperatures. Indicative error bars are obtained by repeating the measurement at several nearby charge degeneracy points. The 2CK (1CK) zone of influence is displayed as a grey (light grey) background. The conductance flows predicted at small $G_{1,2} \ll e^2/h$, for the parameters corresponding to the crossed data line of the same colour, are shown as dashed lines (Methods).

Intriguingly, we also observe (Fig. 4) that the ‘*in situ*’ conductance of the most strongly coupled QPC can slightly overstep the standard quantum limit e^2/h . This overshoot is robust to experimental conditions, above noise level and not a simple calibration artefact (Methods).

The present observation of the two-channel ‘charge’ Kondo effect demonstrates that Kondo physics applies to the degenerate macroscopic quantum states of electrical circuits. Our hybrid device allows full control and characterization of the Kondo parameters, and gives access to ($N \geq 2$)-channel Kondo physics. The implementation in the quantum Hall regime also opens the path to exploring Kondo physics with anyonic quasi-particles, at fractional filling factors. One limitation is the smallness of the charging energy E_C . However, we anticipate that much higher E_C will be feasible by replacing the buried 2DEG with a surface conductor, such as graphene.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 11 March; accepted 28 July 2015.

1. Vojta, M. Impurity quantum phase transitions. *Phil. Mag.* **86**, 1807–1846 (2006).
2. Bulla, R., Costi, T. A. & Pruschke, T. Numerical renormalization group method for quantum impurity systems. *Rev. Mod. Phys.* **80**, 395–450 (2008).

3. Hewson, A. C. *The Kondo Problem to Heavy Fermions* (Cambridge Univ. Press, 1997).
4. Cox, D. L. & Zawadowski, A. Exotic Kondo effects in metals: magnetic ions in a crystalline electric field and tunnelling centres. *Adv. Phys.* **47**, 599–942 (1998).
5. Dzero, M., Sun, K., Galitski, V. & Coleman, P. Topological Kondo insulators. *Phys. Rev. Lett.* **104**, 106408 (2010).
6. Goldhaber-Gordon, D. *et al.* Kondo effect in a single-electron transistor. *Nature* **391**, 156–159 (1998).
7. Cronenwett, S. M., Oosterkamp, T. H. & Kouwenhoven, L. P. A tunable Kondo effect in quantum dots. *Science* **281**, 540–544 (1998).
8. Sasaki, S., Amaha, S., Asakawa, N., Eto, M. & Tarucha, S. Enhanced Kondo effect via tuned orbital degeneracy in a spin 1/2 artificial atom. *Phys. Rev. Lett.* **93**, 017205 (2004).
9. Potok, R. M., Rau, I. G., Shtrikman, H., Oreg, Y. & Goldhaber-Gordon, D. Observation of the two-channel Kondo effect. *Nature* **446**, 167–171 (2007).
10. Kondo, J. Resistance minimum in dilute magnetic alloys. *Prog. Theor. Phys.* **32**, 37–49 (1964).
11. Matveev, K. A. Quantum fluctuations of the charge of a metal particle under the Coulomb blockade conditions. *Sov. Phys. JETP* **72**, 892–899 (1991).
12. Yi, H. & Kane, C. L. Quantum Brownian motion in a periodic potential and the multichannel Kondo problem. *Phys. Rev. B* **57**, R5579–R5582 (1998).
13. Le Hur, K. Kondo resonance of a microwave photon. *Phys. Rev. B* **85**, 140506 (2012).
14. Goldstein, M., Devoret, M. H., Houzet, M. & Glazman, L. I. Inelastic microwave photon scattering off a quantum impurity in a Josephson-junction array. *Phys. Rev. Lett.* **110**, 017002 (2013).
15. Glazman, L. I. & Matveev, K. A. Lifting of the Coulomb blockade of one-electron tunneling by quantum fluctuations. *Sov. Phys. JETP* **71**, 1031–1037 (1990).
16. Matveev, K. A. Coulomb blockade at almost perfect transmission. *Phys. Rev. B* **51**, 1743–1751 (1995).
17. Furusaki, A. & Matveev, K. A. Theory of strong inelastic cotunneling. *Phys. Rev. B* **52**, 16676–16695 (1995).
18. Zaránd, G., Zimányi, G. T. & Wilhelm, F. Two-channel versus infinite-channel Kondo models for the single-electron transistor. *Phys. Rev. B* **62**, 8137–8143 (2000).
19. Le Hur, K. & Seelig, G. Capacitance of a quantum dot from the channel-anisotropic two-channel Kondo model. *Phys. Rev. B* **65**, 165338 (2002).
20. Lebanon, E., Schiller, A. & Anders, F. B. Coulomb blockade in quantum boxes. *Phys. Rev. B* **68**, 041311 (2003).
21. Nozières, P. & Blandin, A. Kondo effect in real metals. *J. Phys.* **41**, 193–211 (1980).
22. Pustilnik, M., Borda, L., Glazman, L. I. & von Delft, J. Quantum phase transition in a two-channel-Kondo quantum dot device. *Phys. Rev. B* **69**, 115316 (2004).
23. Nygård, J., Cobden, D. H. & Lindelof, P. E. Kondo physics in carbon nanotubes. *Nature* **408**, 342–346 (2000).
24. Park, J. *et al.* Coulomb blockade and the Kondo effect in single-atom transistors. *Nature* **417**, 722–725 (2002).
25. Liang, W., Shores, M. P., Bockrath, M., Long, J. R. & Park, H. Kondo resonance in a single-molecule transistor. *Nature* **417**, 725–729 (2002).
26. Jarillo-Herrero, P. *et al.* Orbital Kondo effect in carbon nanotubes. *Nature* **434**, 484–488 (2005).
27. Ralph, D. C., Ludwig, A. W. W., von Delft, J. & Buhrman, R. A. 2-channel Kondo scaling in conductance signals from 2 level tunneling systems. *Phys. Rev. Lett.* **72**, 1064–1067 (1994).
28. Beenakker, C. W. J. Theory of Coulomb-blockade oscillations in the conductance of a quantum dot. *Phys. Rev. B* **44**, 1646–1656 (1991).
29. Joyez, P., Bouchiat, V., Esteve, D., Urbina, C. & Devoret, M. H. Strong tunneling in the single-electron transistor. *Phys. Rev. Lett.* **79**, 1349–1352 (1997).
30. Oreg, Y. & Goldhaber-Gordon, D. Two-channel Kondo effect in a modified single electron transistor. *Phys. Rev. Lett.* **90**, 136602 (2003).
31. Jezouin, S. *et al.* Tomonaga-Luttinger physics in electronic quantum circuits. *Nature Commun.* **4**, 1802 (2013).

Acknowledgements This work was supported by the ERC (ERC-2010-StG-20091028, no. 259033) and the French RENATECH network. We acknowledge E. Boulard, J. von Delft, S. De Franceschi, L. Glazman, D. Goldhaber-Gordon, K. Le Hur, A. Keller, K. Matveev, L. Peeters, P. Simon and G. Zaránd for critical reading of our manuscript and discussions.

Author Contributions Z.I. and F.P. performed the experiment. Z.I., A.A. and F.P. analysed the data. F.D.P. fabricated the sample. U.G. and A.C. grew the 2DEG. S.J. contributed to a preliminary experiment. F.P. led the project and wrote the manuscript with input from Z.I., A.A. and U.G.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to F.P. (frederic.pierre@ipn.cnrs.fr).

METHODS

Experimental setup. The measurements were performed using standard lock-in techniques, at frequencies below 100 Hz, in a dilution refrigerator. Multiple filters along the electrical lines and two shields at the mixing chamber protect the sample from spurious high energy photons.

Sample. The sample is nanostructured by standard e-beam lithography in a 70-nm-deep GaAs/Ga(Al)As 2DEG of density $2.5 \times 10^{11} \text{ cm}^{-2}$ and mobility $10^6 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$. The metallic island is constituted of nickel (30 nm), germanium (60 nm) and gold (120 nm).

Electronic temperature. The electronic temperature T and the associated error bars are obtained from standard quantum shot noise measurements across both QPC_{1,2} and, at $T \geq 38 \text{ mK}$, also from the readings of a RuO₂ thermometer. The temperature stability is ascertained by measuring the electronic temperature before and after data acquisition, as well as with continuous RuO₂ readings. For details on the noise measurement setup see the supplementary materials of ref. 32.

Electronic level spacing in the metallic island. The typical energy spacing between electronic levels in the central metallic island is evaluated from the standard expression $\delta = 1/(v_F \Omega)$, with Ω the island's volume and v_F the electronic density of states per unit volume and energy in the metallic island. Taking the island's volume $\Omega \approx 3 \mu\text{m}^3$ and a typical density of states for metals $v_F \approx 10^{47} \text{ J}^{-1} \text{ m}^{-3}$ (in gold, the main constituent, $v_F \approx 1.14 \times 10^{47} \text{ J}^{-1} \text{ m}^{-3}$), we find $\delta \approx k_B \times 0.2 \text{ mK} \ll k_B T$. The very small electronic level spacing, more than four orders of magnitude smaller than the thermal energy $k_B T$, verifies the essential hypothesis $\delta \ll k_B T$ in the theory^{11,16,17}. To further demonstrate that, in general, the electronic level spacing in the island is fully negligible, one can compare δ with the electronic level energy width \hbar/τ_ϕ , where τ_ϕ is the electronic quantum coherence time. Indeed, $\delta \ll \hbar/\tau_\phi$ corresponds to a continuous electronic density of states. The typical electron quantum coherence time is in the 10 ns range at low temperatures in similarly diffusive metals³³ (see, for example, ref. 34 for the measurement of τ_ϕ in gold). The corresponding electronic level energy width $\hbar/\tau_\phi \approx k_B \times 5 \text{ mK} \gg \delta$ is therefore greater than the typical level spacing by approximately four orders of magnitude.

Interface metallic island and 2DEG. It is crucial to achieve a nearly perfect transmission of the outer electronic channel propagating along the edge of the buried 2DEG towards the central metallic island. Here, we detail the procedure to precisely determine this transmission probability. The notations are recapitulated in Extended Data Fig. 1. In the following, the lateral gates are fully depleted. First, QPC_{1,2,p} are set to the middle of the very flat and large ($\sim 0.4 \text{ V}$) intermediate plateau at $\tau_{1,2,p} = 1$ (thanks to the robust quantum Hall effect, see Extended Data Fig. 2c for the corresponding plateau across a lateral characterization gate) and we measure the corresponding $V_{ii}^{\tau_{1,2,p}=1}$ ($i \in \{1, 2, p\}$). The transmission probability $\tau_{\Omega-i}$ of the outer edge channel from QPC_{*i*} into the metallic island is then given by the expression

$$V_{ii}^{\tau_{1,2,p}=1} = (2 - \tau_{\Omega-i})V_i/2 + \tau_{\Omega-i} \frac{\tau_{\Omega-i}V_i/2}{\tau_{\Omega-1} + \tau_{\Omega-2} + \tau_{\Omega-p}}$$

Note that we made absolutely sure that there were no other ways than through the metallic island to go from QPC_{*i*} to QPC_{*j*}, with $i \neq j$. This is done by etching trenches in the 2DEG underneath the island (see Fig. 1a and Extended Data Fig. 1).

Second, we eliminate calibration uncertainties by measuring the reflected signals $V_{ii}^{\tau_{1,2,p}=0} = V_i$ with QPC_{1,2,p} disconnected (depleted). The ratios $V_{ii}^{\tau_{1,2,p}=1}/V_{ii}^{\tau_{1,2,p}=0}$ give $\tau_{\Omega-i}$ independently of the injection and measurement chain calibrations

$$\frac{V_{ii}^{\tau_{1,2,p}=1}}{V_{ii}^{\tau_{1,2,p}=0}} = (2 - \tau_{\Omega-i})/2 + \tau_{\Omega-i} \frac{\tau_{\Omega-i}/2}{\tau_{\Omega-1} + \tau_{\Omega-2} + \tau_{\Omega-p}} \quad (4)$$

With this approach, we obtain $|1 - \tau_{\Omega-i}| \lesssim 3 \times 10^{-4}$

$$\tau_{\Omega-1} = 0.9997, \quad \tau_{\Omega-2} = 1.0003, \quad \tau_{\Omega-p} = 1.0001$$

The outer edge channel is perfectly transmitted into the metallic island at our experimental accuracy.

Calibration of injection and measurement chains. In the same spirit as above, and now assuming $\tau_{\Omega-i} = 1$, we normalize the signal V_{ij} (see notation in Extended Data Fig. 1) by the signal $V_{ij}^{\tau_{i,j}(k)=1(0)}$ measured when setting $\tau_{i,j} = 1$ with the other QPC disconnected ($\tau_k = 0$). For $i \neq j$, this gives

$$v_{ij} \equiv V_{ij}/V_{ij}^{\tau_{i,j}(k)=1(0)} = \frac{G_i G_j 2\hbar/e^2}{G_1 + G_2 + G_p} \quad (5)$$

The same information can also be extracted by solving the set of three equations for the reflected signals ($i = j$)

$$v_{ii} \equiv V_{ii}/V_{ii}^{\tau_{1,2,p}=0} = (1 - G_i \hbar/2e^2) + \frac{G_i^2 \hbar/2e^2}{G_1 + G_2 + G_p} \quad (6)$$

Note that if $G_p = 0$, the measurements of v_{11} , v_{22} , v_{12} and v_{21} are redundant and only give access to $G_{\text{SET}} = 1/(G_1^{-1} + G_2^{-1})$, but not to G_1 and G_2 separately.

QPC characterization. Extended Data Fig. 2a, b displays as a continuous line the measured 'intrinsic' (not renormalized by Kondo effect or Coulomb blockade) transmission probability $\tau_{1(2)}$ of QPC_{1(2)}} versus the gate voltage $V_{\text{qpc1(2)}}$ applied to one side of the corresponding split gate ($T \approx 11.5 \text{ mK}$, no dc bias voltage). The symbols indicate the QPC set points used in the experiment. Note that for larger (less negative) values of $V_{\text{qpc1,2}}$, the 'intrinsic' QPC conductances exhibit a wide ($\sim 0.4 \text{ V}$) plateau, precisely at e^2/h and robust to dc voltages within the explored range $|V_{\text{dc}}| < 100 \mu\text{V}$. This is followed by a second step up to $2e^2/h$ corresponding to the opening of a second electronic (inner edge) channel (not shown but similar to the lateral characterization gate, see Extended Data Fig. 2c). The insets in Extended Data Fig. 2a, b show the relative variation of the corresponding 'intrinsic' QPC differential conductance with the applied dc bias voltage, up to $|V_{\text{dc}}| = 50 \mu\text{V} \approx 2E_C/e$ and for $\tau_{1,2} \approx \{0.06, 0.47, 0.93\}$ (data shifted vertically by 0.1 for clarity). The relatively small impact of dc bias voltage corroborates a point-like description of the QPCs within the pertinent energy range, below E_C . Note that the broad dip visible in the transmission across QPC₁ at larger split gate voltages V_{qpc1} (Extended Data Fig. 2a) has no impact at the experimental set points used. In particular, it does not result in strongly energy dependent transmission probabilities (inset of Extended Data Fig. 2a, e) and it has no impact on the dynamical Coulomb blockade low bias conductance suppression (Extended Data Fig. 2d). To perform the measurements in Extended Data Fig. 2a, b, both lateral characterization gates (Fig. 1a, coloured yellow for QPC₂, not coloured for QPC₁) were set to zero gate voltage. As shown Extended Data Fig. 2c, this corresponds to fully transmitting the two electronic edge channels across the lateral gates, thereby effectively short-circuiting the central metallic island (in normal operation, the lateral characterization gates are set to about -0.4 V in order to deplete the 2DEG underneath; for further details regarding the lateral characterization gates' 'switch' operation, see the supplementary information in ref. 35). Extended Data Fig. 2d shows as continuous lines the differential conductance across QPC_{1,2} measured at 22 mK as a function of dc voltage with the nearby lateral characterization gate set to deplete the 2DEG (biased at about -0.4 V , as when exploring the 2CK physics) while the lateral gate on the opposite side of the metallic island is set to transmit the two edge channels (biased at $\sim 0 \text{ V}$), as illustrated schematically. The central conductance dip at low dc voltage corresponds to the dynamical Coulomb blockade suppression of the conductance^{31,35}, while the flat plateaus at large dc voltages are used to extract the 'intrinsic' transmission probabilities $\tau_{1,2}$ here displayed as horizontal dashed lines. The precise values of $\tau_{1,2}$ at the experimental set points, and their relative increase $\Delta\tau_{1,2}/\tau_{1,2}$ between zero bias and $\pm 50 \mu\text{V}$ (corresponding to our estimated experimental uncertainty on τ), are recapitulated in Extended Data Fig. 2e.

Capacitive cross-talk. Changing the gate voltage controlling one QPC also slightly affects the others. This cross-talk is determined precisely using the lateral characterization gates, from the shift in gate voltage of the QPC 'intrinsic' conductance curves shown Extended Data Fig. 2a, b. Thanks to the relatively important distances (several micrometres) between QPCs (compared to small quantum dots) the cross-talk correction is small, typically a few per cent. We take into account the small capacitive cross-talk correction during data acquisition.

Charging energy characterization. The charging energy $E_C = e^2/2C \approx k_B \times 290 \text{ mK}$ is obtained from the measured Coulomb diamonds displayed in Extended Data Fig. 3.

Conductance peak reproducibility. Although a single period of $G_{\text{SET}}(V_g)$ is shown Fig. 1c, we systematically measured several nearby periods for each configuration. Extended Data Fig. 4 displays as symbols several consecutive periods measured at base temperature $T \approx 11.5 \text{ mK}$ for the same configuration $\tau = 0.93$ shown in Fig. 1c, together with the quantitative theoretical prediction of equation (9) (continuous line). In practice we take the average of the maximum conductance (at charge degeneracy) measured for different periods, and we estimate the experimental uncertainty from the scatter between values. In Figs 2 and 3, the extracted uncertainty (not shown) is smaller than the symbols. In Fig. 4, the extracted experimental uncertainty is displayed as error bars.

Theoretical expression of G_{SET} and $G_{1,2}$ at $\tau_{1,2} \ll 1$ ($T \gg T_K$). In the limit $T \gg T_K$ (also corresponding to the tunnel regime $\tau_{1,2} \ll 1$), the two Kondo channels are independent from one another since they only weakly screen the pseudospin-1/2. Consequently the 'in situ' (Kondo renormalized) conductances $G_{1,2} \ll e^2/h$ renormalize independently, increasing as temperature is reduced near charge degeneracy ($\delta V_g \approx 0$), owing to the Kondo effect. Theory predicts that the standard expression (equation (1)) for independent sequential tunnelling events holds provided that the 'intrinsic' transmission probabilities $\tau_{1,2}$ in $G_\infty = (e^2/h)/(\tau_1^{-1} + \tau_2^{-1})$ are substituted by the Kondo renormalized values¹⁷

$$\tau_{1,2} \rightarrow \pi^2 / \log^2 \left(\max \{ T, 2E_C |\delta V_g / A| / k_B \} / \alpha T_{K1,2} \right) \quad (7)$$

where α is a numerical factor depending on the precise definition of $T_{K1,2} = T_K(\tau_{1,2})$ (equation (2), see 'Rescaling procedure' below for the determination

of α). Note that the substitution equation (7) leaves the width of the conductance line shape essentially proportional to temperature, although slightly narrower, in reasonable agreement with the data. Consequently, at a good approximation in the tunnel regime, the Kondo effect essentially results in an increased value of the parameter G_∞ in equation (1). In this spirit, we have fitted the tunnel data shown in Fig. 1b (symbols) using equation (1) with the temperature and G_∞ as free parameters (continuous line). In Fig. 2b, at charge degeneracy ($\delta V_g = 0$), the displayed theoretical (thy) $T \gg T_K$ prediction (violet dashed line) is given by

$$G_{\text{SET}}^{\text{thy } T/T_K \gg 1} (T/T_K) = 9.62 \frac{e^2}{h} \log^{-2} \left(\frac{T}{0.0037 T_K} \right) \quad (8)$$

In Fig. 4, the displayed predictions for the renormalization flow at small $G_{1,2} \lesssim 0.6e^2/h$ (dashed lines with the same colour code as the corresponding data) are calculated without additional fit parameters, using $G_{1,2} = 2G_{\text{SET}}^{\text{thy } T/T_K \gg 1} (T/T_{K1,2})$, with $T_{K1,2} = T_K(\tau = \tau_{1,2})$ given by the previously extracted experimental scaling temperature shown in the inset of Fig. 2b, and with $G_{\text{SET}}^{\text{thy } T/T_K \gg 1}$ given by equation (8).

Theoretical expression of G_{SET} at $\tau_{1,2} \approx 1$. The quantitative expression of G_{SET} has been established for arbitrary offsets from the charge degeneracy point (δV_g), in the limit where both QPC_{1,2} are set close to the ballistic limit ($1 - \tau_{1,2} \ll 1$) and for low temperatures with respect to the charging energy $k_B T \ll E_C$ (ref. 17, based on the theoretical framework developed in ref. 16). The prediction shown as a continuous line in Fig. 1c is obtained quantitatively, without fit parameters, from the theoretical expression (equations (38), (26) and (A9) in ref. 17)

$$G_{\text{SET}} = \frac{e^2}{2h} \left[1 - \frac{\pi^2 \gamma \Gamma_+ k_B T}{16 E_C} - \int_0^\infty \frac{\Gamma_-^2 / \cosh^2(x)}{(x \pi^2 k_B T / \gamma E_C)^2 + \Gamma_-^2} dx \right] \quad (9)$$

with $\gamma \approx \exp(0.5772)$ and

$$\Gamma_\pm = 2 - \tau_1 - \tau_2 \pm 2\sqrt{(1 - \tau_1)(1 - \tau_2)} \cos(2\pi \delta V_g / \Delta).$$

Note that we have supplemented equation (38) of ref. 17 with the small correction proportional to $k_B T / E_C$ in equation (A9), following the same procedure used in Fig. 2 of ref. 17. The function Γ_- reduces to zero when the sample is set to display the 2CK effect ($\tau_1 = \tau_2$ and $\delta V_g = 0$). Instead, the integral term with Γ_- in equation (9) determines the crossover from quantum criticality, as further discussed in the next section. In symmetric situations ($\tau_1 = \tau_2$) and at the degeneracy point ($\delta V_g = 0$), all the temperature dependence describing the flow towards the 2CK state (quantum critical point) results from the term proportional to Γ_+ in equation (9). Without additional hypotheses other than $\tau \equiv \tau_1 = \tau_2$ and $\delta V_g = 0$, equation (9) can be reformulated as a universal scaling function, whose value tends linearly towards the quantum critical point $e^2/2h$ when the temperature goes to 0

$$G_{\text{SET}}^{T \ll T_K} (T/T_K^{1-\tau \ll 1}, \delta V_g = 0) = \frac{e^2}{2h} (1 - T/2T_K^{1-\tau \ll 1}) \quad (10)$$

with the Kondo scaling temperature defined as

$$T_K^{1-\tau \ll 1} = \frac{2E_C}{\pi^3 \gamma k_B (1 - \tau)} \quad (11)$$

Note that although for small enough $1 - \tau$ the Kondo temperature can become larger than the charging energy E_C , the latter remains a high energy cutoff for the Kondo physics since at larger energies (for example, $k_B T \gtrsim E_C$) additional charge states of the island become accessible. Note also that equally valid definitions of the Kondo temperature can differ by a constant multiplicative factor: replacing $T/2T_K^{1-\tau \ll 1}$ in equation (10) by $\alpha T/2T_K^{1-\tau \ll 1}$ would change the expression of $T_K^{1-\tau \ll 1}$ by the multiplicative factor α . Here, the definition of $T_K^{1-\tau \ll 1}$ was chosen such that $G_{\text{SET}}^{T \ll T_K} (T = T_K^{1-\tau \ll 1}, \delta V_g = 0) = 0.5e^2/2h$ (although $T = T_K^{1-\tau \ll 1}$ is beyond the range of validity of equation (10)). The red short-dashed line displayed in the main panel of Fig. 2b is the quantitative prediction for $G_{\text{SET}}(T)$ calculated with equation (10) for $\tau = 0.86$ and $E_C = k_B \times 290$ mK. It was rescaled in T/T_K using the same experimental scaling temperature $T_K(\tau = 0.86) \approx 1.4$ K as the $\tau = 0.86$ data (and not $T_K^{1-\tau \ll 1}(\tau = 0.86) \approx 0.075$ K) to allow a direct comparison of data and theory in Fig. 2b. In the inset of Fig. 2b, a constant multiplicative factor is applied to $T_K^{1-\tau \ll 1} \propto 1/(1 - \tau)$ (red short-dashed line) to match the experimental scaling temperature at $\tau \approx 1$.

Predictions for the crossover from quantum criticality. In this section, we show that the predictions of equation (9) correspond to generic expectations for the crossover from quantum criticality^{4,22,36}. An asymmetry between Kondo channels ($\tau_1 - \tau_2 \neq 0$) or a lifting of the charge pseudospin degeneracy ($\delta V_g \neq 0$) is predicted to destroy the unstable 2CK state at vanishing temperatures; and a crossover from non-Fermi liquid (quantum critical) to Fermi liquid behaviour is expected to take place as temperature is reduced. The corresponding crossover temperature is generically expected^{14,22,36} to depend quadratically on the strength of the

perturbations near the symmetric ($\tau_1 = \tau_2$, $\delta V_g = 0$) quantum critical point. The theoretical prediction of equation (9) describes quantitatively the crossover from quantum criticality for the present two-channel ‘charge’ Kondo effect, in the presence of an asymmetry between the two channels and/or of a pseudospin energy splitting. As generically expected, equation (9) predicts that any perturbation ($\tau_1 \neq \tau_2$ and/or $\delta V_g \neq 0$) results in a SET conductance vanishing in the low temperature limit as T^2 , the standard Fermi-liquid power law (see also equation (39) in ref. 17). The crossover behaviour is described by a single function, independent of the perturbations (channel asymmetry $\delta\tau = \tau_1 - \tau_2$, energy splitting $\Delta E = 2E_C \delta V_g / \Delta$, or both simultaneously) that are encapsulated in the parameter Γ_- , thereby corroborating the universal behaviour put forward in ref. 36. The crossover temperature T_{co} can be extracted from equation (9). Here it is defined as the temperature at which $G_{\text{SET}} = 0.5e^2/2h$ (assuming a fully developed 2CK state in absence of perturbation, that is, neglecting the term proportional to $\Gamma_+ T / E_C$ in equation (9)). At $\delta V_g = 0$ and for small $\delta\tau \ll 2 - \tau_1 - \tau_2$, one obtains from equation (9) the crossover temperature $T_{co}(\Delta E = 0, \delta\tau) \approx (\gamma^2 \pi^2 / 4) T_K^{1-\tau \ll 1} (\delta\tau)^2$, corresponding to generic predictions (detailed in, for example, ref. 36). At $\delta\tau = 0$ and for small $\delta V_g \ll \Delta$, one obtains from equation (9) the crossover temperature $T_{co}(\Delta E, \delta\tau = 0) \approx (4/\pi^3) T_K^{1-\tau \ll 1} (\Delta E / k_B T_K^{1-\tau \ll 1})^2$ corresponding to generic predictions³⁶.

Rescaling procedure. We show in Fig. 2b that the data $G_{\text{SET}}(T, \tau)$ for symmetric QPCs ($\tau \approx \tau_1 \approx \tau_2$) can be rescaled into a single curve $G_{\text{SET}}(T/T_K)$. To illustrate the procedure, let us consider two successive transmissions τ and τ' with a conductance overlap such that one can find two data points $G_{\text{SET}}(T, \tau) = G_{\text{SET}}(T', \tau')$. The existence of a universal scaling law directly implies $T_K(\tau')/T_K(\tau) = T'/T$. If such a law exists, then the rescaled data at τ and τ' should match on the full range of conductance overlap. This scheme does not apply directly for the three lowest transmission probabilities $\tau \approx \{0.06, 0.125, 0.245\}$, since there is no conductance overlap. However, theory predicts¹⁷ in the corresponding limit $T \gg T_K$ that $G_{\text{SET}}(T) \propto \log^{-2}(T/\alpha T_K)$. Using this expression to fit and extrapolate the $\tau < 0.25$ data points (dashed line in Fig. 2b), we can apply the above procedure. Note that $T_K(\tau)$ is extracted only up to an overall prefactor. Following standard usage³⁷, we set this prefactor such that $G_{\text{SET}}(T/T_K = 1) = 0.5e^2/2h$ (half the 2CK state conductance).

Absence of numerical renormalization group calculations for $G_{\text{SET}}(T/T_K)$. To the best of our knowledge, there are no available numerical calculations for the measured conductance G_{SET} . Consequently, there is no theoretical prediction to compare with the experimentally extracted scaling curve $G_{\text{SET}}(T/T_K)$ shown in Fig. 2b, beyond the limits of large or small T/T_K . Quoting the authors of ref. 22, the root of the difficulty “is that there is no mapping between the conductance across the island (G_{SET}) and the electron scattering cross-section in the generic two-channel Kondo model”. Hopefully, future numerical work, adapted to the present charge Kondo implementation, will fill this gap and allow a full quantitative comparison of data and theory, including intermediate values of T/T_K .

Extracting separately the ‘in situ’ conductances G_1 and G_2 with QPC_p. The SET conductance, with QPC_p disconnected, only gives access to the series combination of the ‘in situ’ (Kondo renormalized) conductances G_1 and G_2 ($G_{\text{SET}} = 1/(G_1^{-1} + G_2^{-1})$). This is sufficient in symmetric configurations $G_1 \approx G_2$, but not to extract the full renormalization flow shown in Fig. 4. For this purpose we use an additional probe QPC_p. To minimize the effect of this probe, it is set to a relatively small coupling with respect to G_1 and G_2 . In practice, $1/150 < G_p/\min(G_1, G_2) < 1/6$, with the largest values corresponding to the most asymmetric configurations between QPC_{1,2}. As easily checked from equation (5), this gives access directly to $G_1/G_2 = v_{p1}/v_{p2}$ (or equivalently, $G_1/G_2 = v_{1p}/v_{2p}$). Solving equations (5) and (6), with the measured v_{ij} , gives all three ‘in situ’ conductances $G_{1,2,p}$ (provided that $G_{1,2,p} \neq 0$).

‘In situ’ conductances above the standard quantum limit e^2/h . Some of the ‘in situ’ conductances displayed in Fig. 4 slightly overstep the standard quantum limit e^2/h for asymmetric QPC configurations and at low temperatures. Although the standard quantum limit applies to a single quantum channel connected to voltage biased reservoirs (in contrast, the central metallic island is floating) and in the absence of interactions, to the best of our knowledge such behaviour has not previously been observed. In principle, a partial transmission of the second (inner) edge channel across the QPC could provide a simple explanation for the observation of an ‘in situ’ conductance above e^2/h . However this is unlikely because the second electronic channel is initially completely reflected, separated from the full opening of the first (outer) channel by a plateau that is very wide and very robust to dc voltage (tested up to $|V_{dc}| \approx 100 \mu\text{V} \approx 4E_C/e$). Note that we checked ‘in situ’ that the lateral characterization gates set to reflect the two edge channels ($\tau_{lbg} = 0$, at $V_{lbg} \approx -0.4$ V), as well as QPC_p when initially disconnected, remain in this configuration in the presence of the charge Kondo effect. It is also noteworthy that in the present experimental configuration, in the integer quantum Hall regime at filling factor $\nu = 2$, the current

between the metallic island and the QPCs is carried by two co-propagating quantum Hall channels that are coupled by the Coulomb interaction. However, for the short distance between island and QPC and the very low temperatures in the present experimental investigation, this coupling is expected to be negligible³⁸. A similar transient overshoot is predicted in the related Luttinger liquid problem (at $K < 1/2$, see figure 1 in ref. 39), which corresponds to an ‘*in situ*’ single channel differential conductance above e^2/h (in the context of the Luttinger liquid-dynamical Coulomb blockade mapping^{31,40}).

We show here that our intriguing observation is well above the noise level, that the same result is obtained with different sets of measurements, and that it is robust with respect to injection voltage and to the coupling of QPC_p. For this purpose we focus on the set point ($\tau_1 = 0.76$, $\tau_2 = 0.93$) at $T \approx 14$ mK. Extended Data Fig. 5a–d shows the normalized transmitted signals v_{ij} with $i \neq j$ and the reflected signal $2 - v_{pp}$, measured in the linear response regime here with $V_i \approx 1.15 \mu\text{V}_{\text{rms}} < k_B T/e$. The displayed statistical error bars are obtained by repeating the measurements ten times in a row for each data point. Possible charge offset jumps are ruled out from the reproducibility of the two displayed consecutive sweeps (V_g increasing and decreasing). Note that the same normalized data are found for the reciprocal signals $v_{ij} \approx v_{ji}$. Note also that QPC_p is here set to a different tuning (with a higher conductance) than the corresponding data point in Fig. 4. First, we extract $G_{1,2,p}$ solving equation (5) with only the transmitted signals (v_{ij} with $j \neq i$). Averaging all the data (V_g increasing and decreasing, and reciprocal signals) at the degeneracy point, we find

$$G_1 = 0.508 \pm 0.003 e^2/h$$

$$G_2 = 1.11 \pm 0.02 e^2/h$$

$$G_p = 0.0387 \pm 0.0005 e^2/h$$

Second, we show that the same result is obtained with a different set of measurements, now involving also equation (6). We use $v_{12}/2$ and $2(1 - v_{pp})$ (Extended Data Fig. 5d) corresponding to $1/(G_1^{-1} + G_2^{-1})$ and G_p , respectively, in the limit

$G_p \ll G_{1,2}$, as well as $v_{1p}/v_{2p} = G_1/G_2$ (Extended Data Fig. 5e). This gives the consistent values

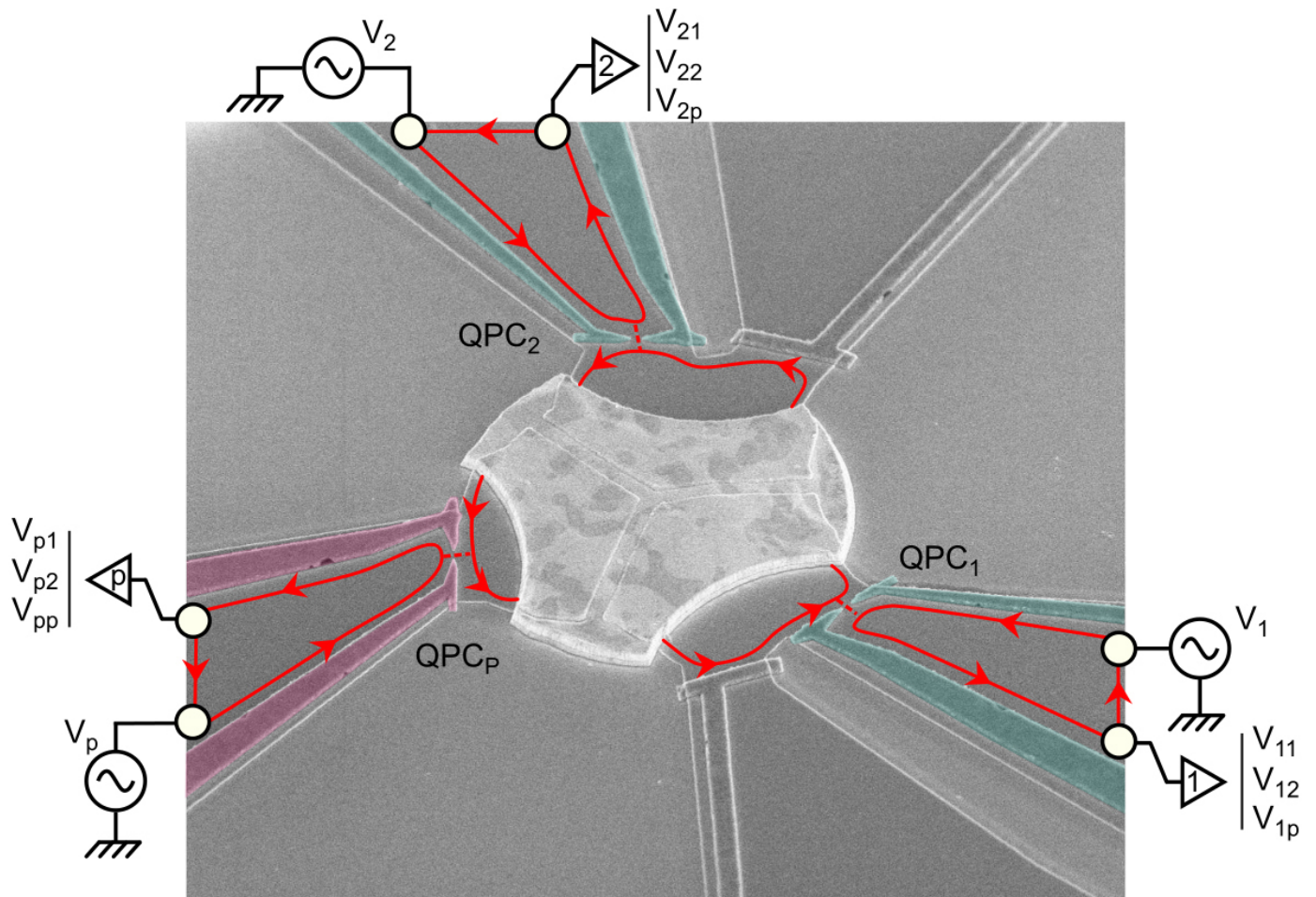
$$G_1 = 0.510 \pm 0.002 e^2/h$$

$$G_2 = 1.107 \pm 0.006 e^2/h$$

$$G_p = 0.0395 \pm 0.0002 e^2/h$$

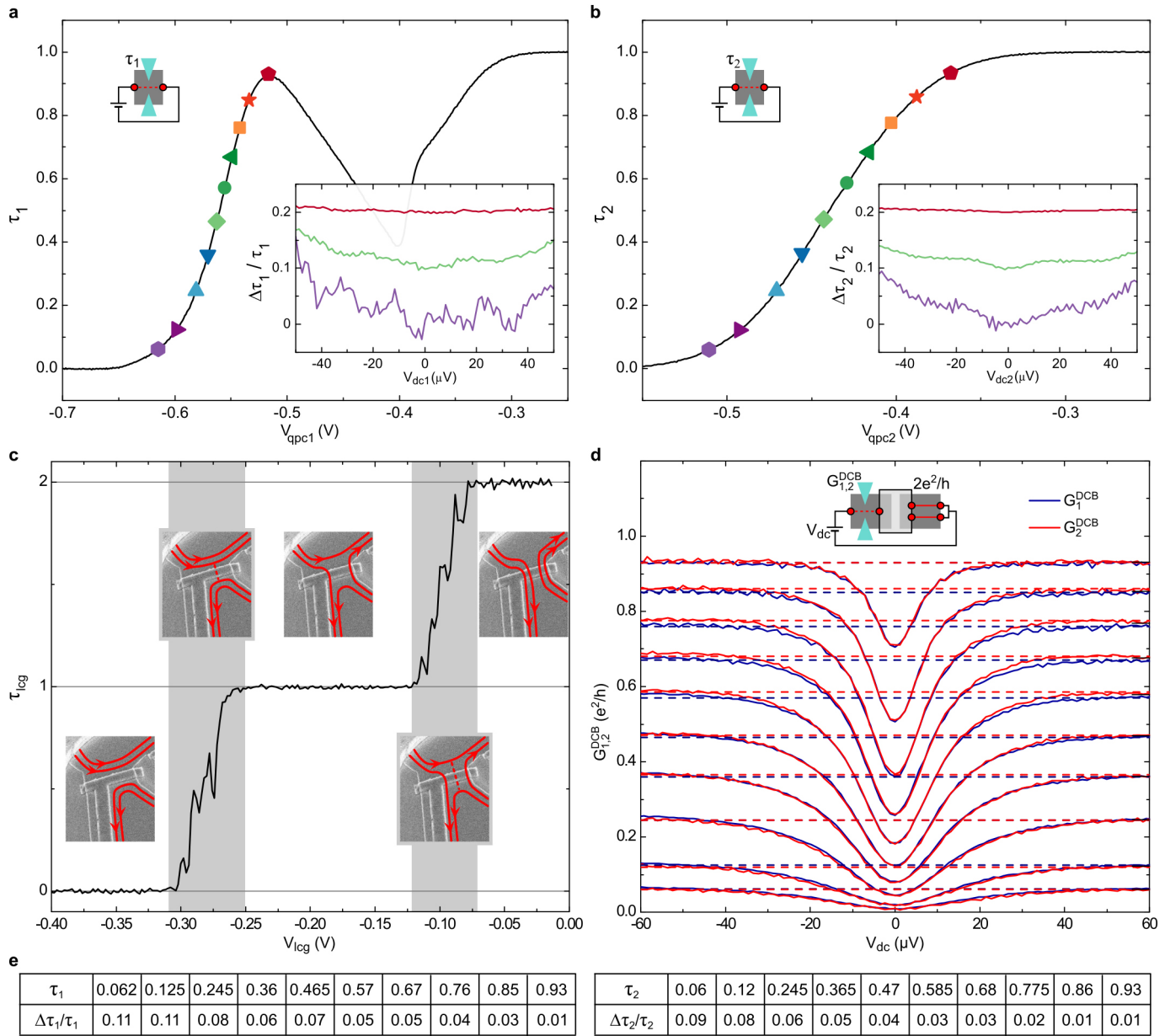
In fact, no averaging is required to show that $G_2 > e^2/h$ well beyond the noise level. Focusing on the smallest signal $v_{1p}/v_{2p} = G_1/G_2$, we observe directly in Extended Data Fig. 5e that every data point near $\delta V_g \approx 0$ is below the red line displaying the ratio for which $G_2 = e^2/h$. Every data point therefore corresponds to $G_2 > e^2/h$. To further confirm $G_2 > e^2/h$, we have also checked that this observation is robust with respect to injection voltage V_i and to the value of G_p , as shown Extended Data Fig. 6a, b.

32. Jezouin, S. *et al.* Quantum limit of heat flow across a single electronic channel. *Science* **342**, 601–604 (2013).
33. Pierre, F. *et al.* Dephasing of electrons in mesoscopic metal wires. *Phys. Rev. B* **68**, 085413 (2003).
34. Wellstood, F. C., Urbina, C. & Clarke, J. Hot-electron effects in metals. *Phys. Rev. B* **49**, 5942–5955 (1994).
35. Parmentier, F. D. *et al.* Strong back-action of a linear circuit on a single electronic quantum channel. *Nature Phys.* **7**, 935–938 (2011).
36. Sela, E., Mitchell, A. K. & Fritz, L. Exact crossover Green function in the two-channel and two-impurity Kondo models. *Phys. Rev. Lett.* **106**, 147202 (2011).
37. Goldhaber-Gordon, D. *et al.* From the Kondo regime to the mixed-valence regime in a single-electron transistor. *Phys. Rev. Lett.* **81**, 5225–5228 (1998).
38. le Sueur, H. *et al.* Energy relaxation in the integer quantum Hall regime. *Phys. Rev. Lett.* **105**, 056803 (2010).
39. Fendley, P., Ludwig, A. W. W. & Saleur, H. Exact nonequilibrium transport through point contacts in quantum wires and fractional quantum Hall devices. *Phys. Rev. B* **52**, 8934–8950 (1995).
40. Safi, I. & Saleur, H. One-channel conductor in an ohmic environment: mapping to a Tomonaga-Luttinger liquid and full counting statistics. *Phys. Rev. Lett.* **93**, 126602 (2004).



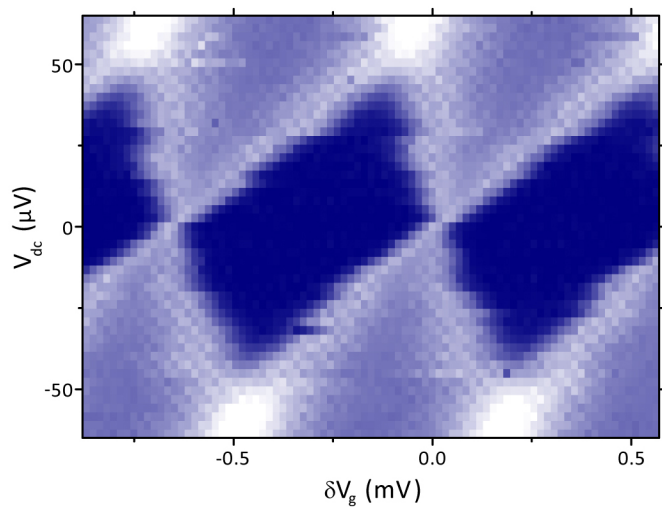
Extended Data Figure 1 | Measurement schematic. Schematic of the measurement setup, showing explicitly the nine different and simultaneously measured signals. V_{ij} ($i, j \in \{1, 2, p\}$) is the voltage measured with amplification

chain i in response to the injected voltage V_j . Trenches etched in the 2DEG in the form of a Y can be seen through the metallic island.

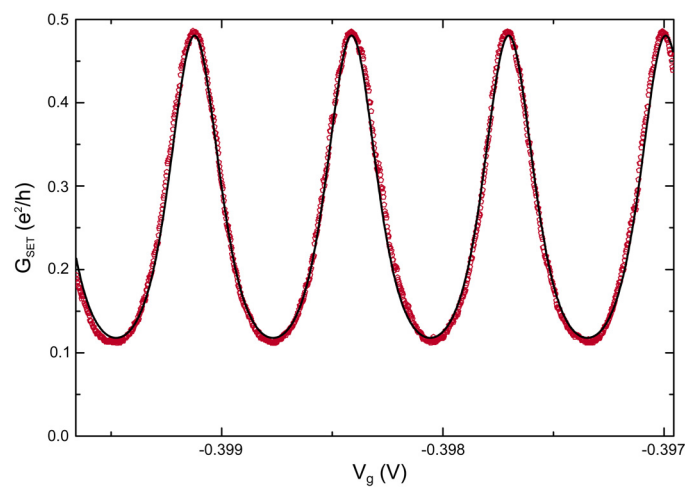


Extended Data Figure 2 | QPC characterization. **a, b**, Main panels, ‘intrinsic’ transmission probability across QPC₁ (τ_1 ; **a**) and QPC₂ (τ_2 ; **b**) measured at 11.5 mK (in the linear regime, without dc bias) by opening the QPC lateral characterization gate (see equivalent schematic in top left insets), and plotted versus the voltage applied to the split gate tuning the QPC ($V_{\text{qpc1,2}}$). The experimental transmission set points in the main text are indicated by symbols. Right insets, relative variation of the transmission probability with dc bias voltage, shifted vertically for clarity, for $\tau_{1,2} \approx \{0.06, 0.47, 0.93\}$ from bottom to top, respectively. The larger noise in the inset of **a** (mostly visible for $\tau_1 \approx 0.06$) is from the amplification chain. **c**, ‘Intrinsic’ conductance across one lateral characterization gate in units of e^2/h (τ_{icg} , here adjacent to QPC1) plotted versus lateral gate voltage V_{icg} . Increasing V_{icg} results in the successive full opening of two electronic channels, as schematically illustrated. In practice, we close (open) the lateral characterization gates, corresponding to $\tau_{\text{icg}} = 0$ ($\tau_{\text{icg}} = 2$), by

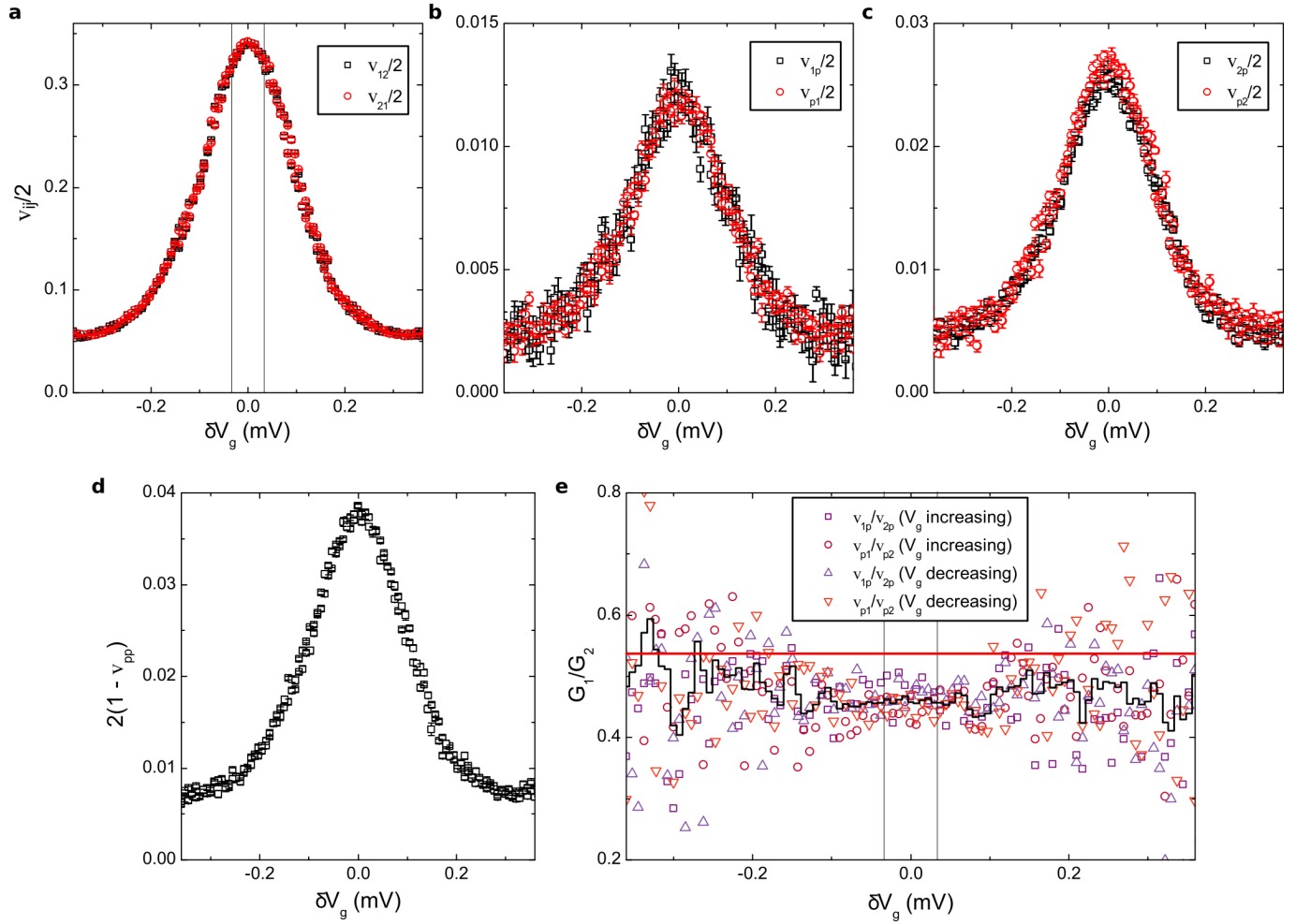
applying $V_{\text{icg}} \approx -0.4$ V ($V_{\text{icg}} = 0$ V). Grey shaded areas correspond to the partial opening of one of the channels, a configuration not used in the experiment. **d**, Conductance of the QPCs measured at $T = 22$ mK versus dc voltage (continuous lines) with the adjacent lateral gate closed ($\tau_{\text{icg}} = 0$) and the lateral characterization gate opposite to the metallic island set to full transmission ($\tau_{\text{icg}} = 2$), which corresponds to the displayed schematic circuit. The low bias dips result from conductance suppression by the dynamical Coulomb blockade, while the high bias plateaus correspond to the ‘intrinsic’ transmission probabilities $\tau_{1,2}$ (horizontal dashed lines). **e**, ‘Intrinsic’ transmission probabilities $\tau_{1,2}$ at the experimental set points used in the main text, together with their relative increase $\Delta\tau_{1,2}/\tau_{1,2}$ between $V_{\text{dc1,2}} = 0$ and $|V_{\text{dc1,2}}| = 50$ μV . This increase is the main experimental factor of uncertainty in the determination of $\tau_{1,2}$.



Extended Data Figure 3 | Coulomb diamonds. The conductance G_{SET} (colour coded; brighter for larger G_{SET}) is displayed versus gate and dc voltages (δV_g and V_{dc} , respectively), with both QPCs set to a low transmission probability. The Coulomb diamonds (darker) correspond to a charging energy $E_C = e^2/2C \approx k_B \times 290 \text{ mK}$.

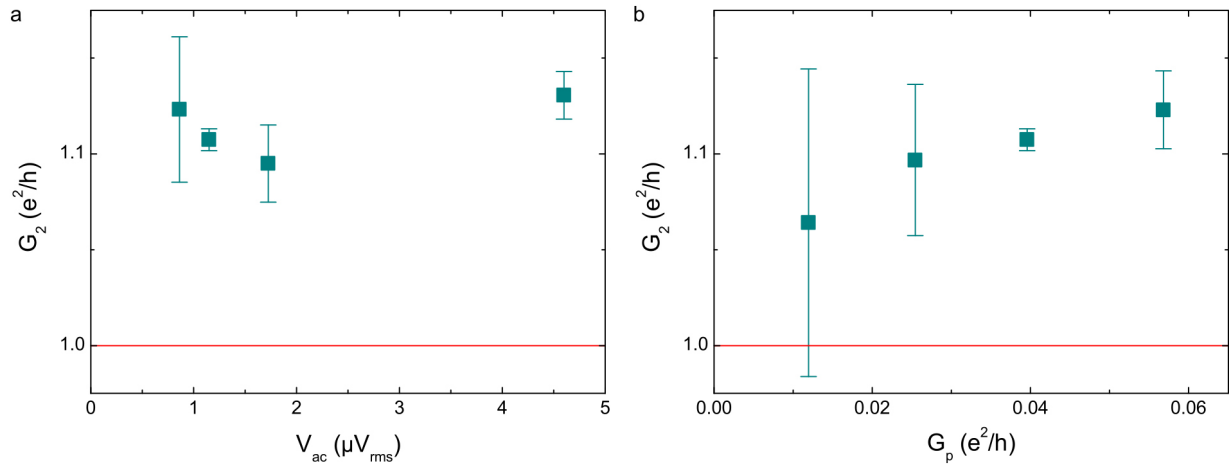


Extended Data Figure 4 | Reproducibility of conductance oscillations. Shown is conductance across the hybrid SET (G_{SET}) when sweeping the gate voltage (V_g) across several periods for the symmetric configuration $\tau_1 \approx \tau_2 \approx 0.93$ and at base temperature $T \approx 11.5$ mK (one period shown in Fig. 1c). The symbols display the measurements, the continuous line is the quantitative prediction of equation (9) in Methods.



Extended Data Figure 5 | Observation of ‘*in situ*’ conductances that overstep the standard quantum limit e^2/h . The displayed Coulomb peaks were measured at $T \approx 14$ mK for the asymmetric QPCs configuration ($\tau_1 \approx 0.77$, $\tau_2 \approx 0.93$). Two sweeps (V_g increasing and decreasing) are shown for each measurement. **a–c**, Symbols are the normalized transmitted signal $v_{ij}/2$, with $i \neq j$ (equation (5) in Methods) versus gate voltage. Each panel displays the two reciprocal signals $v_{i,j}$ and $v_{j,i}$. The vertical lines in **a** are visual markers used in **e**. **d**, Symbols are the normalized reflected signal at the

probe QPC_p ($2(1 - v_{pp})$), corresponding to $G_p h/e^2$ in the limit $G_p \ll G_{1,2}$. **e**, Symbols are the *in situ* conductance ratio G_1/G_2 , measured from both v_{1p}/v_{2p} and v_{p1}/v_{p2} . For each measurement, the two sweeps (V_g increasing and decreasing) are shown with different symbols. The black line is an average at a given δV_g . The red line shows the value below which $G_2 > e^2/h$ near charge degeneracy ($\delta V_g \approx 0$). The error bars shown in **a–d** represent the statistical uncertainties (s.e.m.) calculated from 10 successive measurements.



Extended Data Figure 6 | Robustness to experimental conditions of ‘*in situ*’ conductances overstepping e^2/h . Symbols display the ‘*in situ*’ conductance G_2 measured at $T \approx 14$ mK for the QPC setting $\tau_1 \approx 0.77$ and $\tau_2 \approx 0.93$, and under different experimental conditions. We repeatedly find $G_2 > e^2/h$. **a**, The influence on G_2 of the ac injection voltages $V_{1,2,p}$. The three lowest V_{ac}

correspond to $V_1 = V_2 = V_p = V_{ac}$, whereas the fourth data point corresponds to $V_p = V_{ac}$ with $V_1 = V_2 = 1.15 \mu V_{rms}$. **b**, Exploration of the influence on G_2 of the coupling strength of QPC_p, characterized by G_p . The error bars represent the statistical uncertainties (s.e.m.) calculated from 20 or more different measurements.

Universal Fermi liquid crossover and quantum criticality in a mesoscopic system

A. J. Keller^{1†}, L. Peeters¹, C. P. Moca^{2,3}, I. Weymann⁴, D. Mahalu⁵, V. Umansky⁵, G. Zaránd² & D. Goldhaber-Gordon¹

Quantum critical systems derive their finite-temperature properties from the influence of a zero-temperature quantum phase transition¹. The paradigm is essential for understanding unconventional high- T_c superconductors and the non-Fermi liquid properties of heavy fermion compounds. However, the microscopic origins of quantum phase transitions in complex materials are often debated. Here we demonstrate experimentally, with support from numerical renormalization group calculations, a universal crossover from quantum critical non-Fermi liquid behaviour to distinct Fermi liquid ground states in a highly controllable quantum dot device. Our device realizes the non-Fermi liquid two-channel Kondo state^{2,3}, based on a spin-1/2 impurity exchange-coupled equally to two independent electronic reservoirs⁴. On detuning the exchange couplings we observe the Fermi liquid scale T^* , at energies below which the spin is screened conventionally by the more strongly coupled channel. We extract a quadratic dependence of T^* on gate voltage close to criticality, and validate an asymptotically exact description of the universal crossover between strongly correlated non-Fermi liquid and Fermi liquid states^{5,6}.

A conventional second-order quantum phase transition (QPT) features quantum mechanical fluctuations of a classical order parameter. Some second-order QPTs in heavy fermion materials, notably $\text{CeCu}_{6-x}\text{Au}_x$ and YbRh_2Si_2 , defy easy description in this scheme, and their quantum critical behaviour instead appears to be related to the breakdown of Kondo screening⁷. Distinctive non-Fermi liquid behaviours appear above a so-called Fermi liquid (FL) scale that vanishes at the quantum critical point (QCP); away from the QCP, a crossover from non-FL to FL behaviour is observed at low energies. A diverging effective mass m^* at the QCP, seen in both materials, signifies the absence of quasiparticles at the Fermi surface⁸.

In many heavy fermion materials and in high- T_c superconductors, the relevant degrees of freedom and the effective Hamiltonian can be controversial. We aim to understand quantitatively a second-order QPT outside the usual order-parameter-fluctuation description. Quantum dots provide an experimental framework for realizing known quantum impurity Hamiltonians that can feature tunable second-order QPTs^{9,10}. However, QCPs are challenging to reach even in engineered systems, since perturbations that steer away from quantum criticality may be inherently uncontrolled, as in two-impurity Kondo experiments to date^{11–13}.

At the QCP of a two-channel Kondo (2CK) system, a single overscreened spin yields a non-FL state with no quasiparticles (that is, only collective excitations) at the Fermi surface. The overscreened spin gives rise to a decoupled Majorana mode at the impurity site, to which the non-FL behaviour has been attributed¹⁴. An order parameter is typically not invoked. A FL scale T^* results from several relevant perturbations: Zeeman splitting, difference in exchange couplings and charge transfer between the two channels. Requiring that all these perturbations be small would seem to diminish prospects for observing the QCP in bulk systems. Nonetheless, 2CK physics has been invoked to explain experiments on heavy fermion materials^{15–17} and two-level tunnelling centres^{18–20}.

A 2CK state has been predicted² and observed³ in a quantum dot tunnel-coupled to leads and to a ‘metallic grain’, an electron reservoir big enough to have a small, ideally metallic level spacing $\Delta \lesssim kT$, but small enough to retain a charging energy $E_C \gg kT$ at temperatures of interest. The leads and the grain serve as two independent screening channels for the dot spin. Simply having two conventional leads coupled to the dot is not enough to realize the 2CK state, as charge transfer between the leads cannot be avoided. In contrast, the grain’s

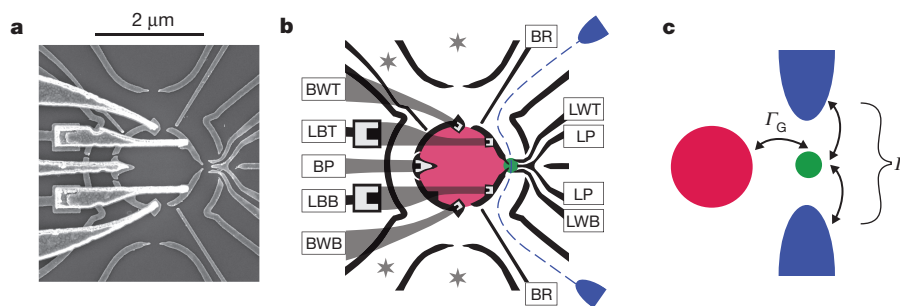


Figure 1 | Device and model. **a**, SEM micrograph of a device nominally identical to the device studied. The five brighter features seen coming in from the left are metal bridges suspended above the sample surface. **b**, Schematic of the device with labelled gate electrodes. Gates BWT, BP and BWT define the grain (red) along with LBT and LBB; the last two also control the dot–grain coupling. Gates LWT, LP and LWB define the dot (green), along with LBT and LBB. Gates BR are used to isolate the dot measurement circuit. Other gates are held at a fixed voltage throughout the experiment. Conductance is

measured between source and drain leads (blue). The four grey stars indicate additional ohmic contacts which are floated during measurement. **c**, Model of the system used for the NRG calculations. Γ_G is the dot–grain coupling; Γ is the total dot–lead coupling (sum of couplings to source and drain leads, Γ_S and Γ_D , respectively). The source and drain leads together act as one channel in the spin 2CK regime, and the Coulomb-blockaded grain acts as an independent channel. The full Hamiltonian is given in Methods section ‘Hamiltonian’.

¹Geballe Laboratory for Advanced Materials, Stanford University, Stanford, California 94305, USA. ²BME-MTA Exotic Quantum Phases “Lendület” Group, Institute of Physics, Budapest University of Technology and Economics, H-1521 Budapest, Hungary. ³Department of Physics, University of Oradea, Oradea 410087, Romania. ⁴Faculty of Physics, Adam Mickiewicz University, Poznań 61-614, Poland. ⁵Department of Condensed Matter Physics, Weizmann Institute of Science, Rehovot 96100, Israel. [†]Present address: Institute for Quantum Information and Matter, California Institute of Technology, Pasadena, California 91125, USA.

charging energy strongly suppresses charge transfer with the other leads. Spin exchange with the dot remains possible, so the grain and leads compete and overscreen the dot spin. In ref. 3, the resulting non-FL behaviour was observed, as were conventional FL single-channel Kondo states far from the QCP. However, T^* was not identified and the non-FL to FL crossover was not seen. In particular, how the FL scale T^* vanishes on approach to the QCP is an important hallmark of the phase transition, with associated critical exponents. Recently, in addition to prior numerical²¹ and analytical²² descriptions of the crossover, a new description has been found using Abelian bosonization and conformal field theory (CFT) methods, yielding asymptotically exact predictions for conductance in the regime where eV_{SD} , kT , $kT^* \ll kT_K$, with T_K the 2CK temperature and V_{SD} the bias voltage on a weakly-coupled probe^{5,6}. This description has the decoupled Majorana mode at the impurity site coupling to degrees of freedom in the reservoirs for non-zero T^* .

In this work, we show how fine control over the 2CK state in a mesoscopic device allows direct comparison to exact results in the crossover regime, yielding T^* as a function of gate detuning away from the QCP. The device (Fig. 1a) is fabricated by lithographically patterning gate electrodes on a GaAs/Al_{0.3}Ga_{0.7}As heterostructure hosting a two-dimensional electron gas (2DEG). A plan of the device is given in Fig. 1b. Despite the number of gates, the device is conceptually simple (Fig. 1c): a metallic grain (red) and two leads (blue) are each tunnel-coupled to a quantum dot (green) at rates Γ_G and Γ , respectively. The total dot-lead tunnel rate $\Gamma = \Gamma_S + \Gamma_D$, the sum of tunnel rates to the source and drain leads. The charging energy is U (E_C) for the dot (grain), and the full Hamiltonian is given in Methods section ‘Hamiltonian’. In this experiment, two-terminal conductance $G = dI/dV_{SD}$ is measured between the pair of leads (Methods section ‘Measurements’). We use gate voltage V_{BWT} (V_{LP}) to tune the grain level ϕ (dot level ϵ) (gates are named in Fig. 1b).

We first identify the set of QCPs in the $(-\epsilon/U, -\phi/E_C)$ plane for fixed Γ , Γ_G . For our model Hamiltonian, quantum critical ‘2CK lines’ periodic in the grain charge are expected instead of isolated QCPs^{2,23,24}. Figure 2a shows the 2CK lines overlaid on numerical renormalization group (NRG) calculations of $G(-\epsilon/U, -\phi/E_C)$ using realistic device parameters. We focus on the spin 2CK regime, though charge fluctuations may be important elsewhere²⁵. To directly compare to the experimentally measured conductance data of Fig. 2c, Fig. 2b adjusts the NRG calculations of Fig. 2a to account for the cross-capacitance between V_{LP} and the grain.

To identify transport signatures of quantum criticality along the 2CK line, we look for the characteristic square-root scaling of $G(V_{SD}, T)$ derived from the CFT of ref. 26. The CFT yields temperature-dependent spectral functions $A_{2CK}(\omega, T, \delta_p)$, where δ_p is a phase shift from potential scattering. These are closely related to $G(V_{SD}, T)$ for $\hbar\omega \rightarrow -eV_{SD}$ (Methods section ‘Relationship of $G(V_{SD}, T)$ to spectral functions’). A scaling collapse of $G(V_{SD}, T)$ is expected:

$$\frac{G(0, T) - G(V_{SD}, T)}{\sqrt{T}} \propto \frac{1}{\sqrt{T_K}} Y_{2CK}(-eV_{SD}/kT, \delta_p) \quad (1)$$

where T_K (Kondo temperature) is a scale below which the 2CK physics is observed and $Y_{2CK}(-eV_{SD}/kT, \delta_p)$ a universal function closely related to $A_{2CK}(\omega, T, \delta_p)$ (Methods section ‘Fitting expressions for 2CK’).

Figure 2d shows spectral functions $A(\omega, T)$ calculated by NRG. Importantly, the spectral functions collapse onto $A_{2CK}(\omega, T, \delta_p)$, with the horizontal axis scaled to emphasize the $\omega^{1/2}$ behaviour for large ω/T . Measured $G(V_{SD}, T)$ on or very near the 2CK line (Fig. 2e) collapse similarly, except for the 20 mK and 40 mK traces at positive V_{SD} . This deviation could result from a small $T^* \lesssim T_e$, the base electron temperature. Data taken at more negative V_{LP} show very clear 2CK scaling (Methods section ‘Scaling along the quantum critical lines’) but are less suitable for analysing the crossover. Experimental $kT_K \approx 50 \mu\text{eV}$ should only be trusted up to factors of order unity: in equation (1), T_K enters only as a scale factor, and other scale factors like

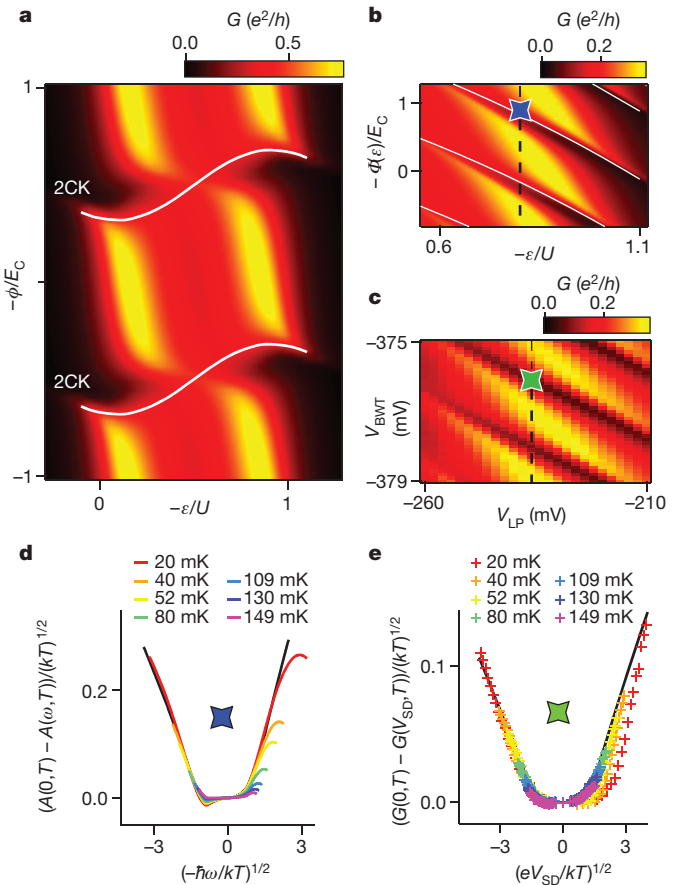


Figure 2 | Quantum phase transitions. **a**, NRG calculations of $G(-\phi/E_C, -\epsilon/U)$ for symmetric source-drain coupling ($V_{SD} = 0$, $T = 20$ mK). Parameters as follows: $U = 2$ meV, $\Gamma = 0.123$ meV, $\Gamma_G = 0.106$ meV, $E_C = 0.15$ meV, bandwidth $D = 1$ meV. 2CK lines (white) are determined by analysis of the finite size spectrum. **b**, The calculations of **a** plotted with an ϵ -dependent shift in ϕ and rescaled by a constant factor for comparison with **c**, to account for unequal source-drain couplings. 2CK lines are shown white. The dashed line indicates $-\epsilon/U = 0.8$, the cut direction of Fig. 3e. **c**, Experimentally measured $G(V_{BWT}, V_{LP})$ ($V_{SD} = 0$, $T = 20$ mK). Gates V_{BWT} and V_{LP} act approximately like $-\phi$ and $-\epsilon$. The dashed line indicates $V_{LP} = -236$ mV, the cut direction of Fig. 3d. **d**, NRG calculations of the equilibrium spectral functions $A(\omega, T)$ for ϵ, ϕ as marked in **b** by the blue star. The black trace is the spectral function $A_{2CK}(\omega, T, \delta_p)$ from CFT ($\delta_p = -0.029\pi$, $kT_K = 19 \mu\text{eV}$). **e**, Measured $G(V_{SD}, T)$ for V_{LP}, V_{BWT} as marked in **c** by the green star. The black trace is $Y_{2CK}(-eV_{SD}/kT, \delta_p)/\sqrt{kT_K}$, rescaled on the basis of an estimate of source-drain coupling asymmetry ($\delta_p = -0.016\pi$, $kT_K = 50 \mu\text{eV}$). The range in (eV_{SD}/kT) decreases as temperature increases because we measure a fixed range in V_{SD} .

source-drain coupling asymmetry must be estimated. We estimate that $\Gamma_S/\Gamma_D \approx 0.15$, and consider the source lead weakly coupled (Methods section ‘Measurements’).

Having identified the 2CK lines in Fig. 2, we consider how to perturb the quantum critical state. In the 2CK model, a single FL scale T^* suffices to describe any combination of symmetry-breaking perturbations^{5,21}. The limit $\hbar\omega$, kT , $kT^* \ll kT_K$ permits an exact expression for the scattering \mathcal{T} matrix in the low-temperature 2CK crossover^{5,6}. In our experimental configuration, the \mathcal{T} matrix is diagonal:

$$2\pi i v T_{\sigma\alpha, \sigma\alpha}(\omega, T, \delta_p) = 1 - e^{2i\delta_p} S_{\sigma\alpha, \sigma\alpha} G\left(\frac{\hbar\omega}{kT^*}, \frac{T}{T^*}\right) \quad (2)$$

with the universal complex-valued function $G(\hbar\omega/kT^*, T/T^*)$ encoding the crossover physics. These diagonal elements relate to $A(\omega, T)$ and thus to experimental $G = dI/dV_{SD}$ for highly asymmetric source-drain coupling. v is the bare density of states per spin in the leads, σ is the spin

index, and $\alpha = 1$ (-1) labels electrons in the leads (grain). The S matrix gives a (spin and channel dependent) scattering phase shift that is a function of the relative strengths of any perturbations present. Negligible charge transfer between channels and zero magnetic field yields $S_{\sigma\alpha, \sigma\alpha} = \pm\alpha$, with $+$ ($-$) indicating the dot is more strongly exchange-coupled to the grain (leads). The factor $e^{2i\delta_P}$ accounts for additional spin-independent phase shifts from potential scattering. We fix $S_{\sigma\alpha, \sigma\alpha} = \alpha$ and let δ_P jump by $\pi/2$ to account for sign changes.

To observe the FL crossover experimentally, we fix $V_{LP} = -236$ mV (dashed line in Fig. 2c) and detune the exchange couplings using V_{BWT} . Moving slightly away from the QCP so that $T^* \approx T_c$, we still measure a $T^{1/2}$ scaling collapse for $T > 50$ mK (Fig. 3a). These high- T data are fitted nicely using the ref. 26 CFT result with small δ_P (black line). The clear scaling behaviour at high T can only be observed for V_{BWT} in a small neighbourhood around the QCP. Below 50 mK, prominent deviations from 2CK scaling develop, which we attribute to a crossover into a FL state where the grain screens the dot spin. Near zero bias these low- T traces are fitted by the crossover theory with similar, small δ_P (Fig. 3b). We stress this is a non-trivial regime since $T^* \approx T_c$; asymptotics of the FL fixed point are insufficient to describe the observed behaviour. For larger $|eV_{SD}/kT|^{1/2}$, the $|eV_{SD}|^{1/2}$ dependence of $G(V_{SD})$, appearing linear on these axes, heralds a return to 2CK behaviour.

Near the QCP, T^* should depend quadratically on the strength of symmetry-breaking perturbations^{5,6,22} (Fig. 3c). At $B = 0$, we expect $T^* \propto |J_1 - J_2|^{2\nu}$, where $J_{1(2)}$ is the exchange coupling for channel 1(2), and the critical exponent $2\nu = 2$ (see ref. 1 for definition of ν and ν). As is generically true for QPTs, the constant of proportionality is non-universal and may differ on either side of the QCP¹. Measured $G(V_{SD}, V_{BWT})$ reveals periodic zero bias dips that transition sharply to zero bias peaks as V_{BWT} is increased (Fig. 3d, top). The zero bias dip (peak) corresponds to a $T = 0$ ground state where the grain (lead) screens the dot spin; these are separated by a QCP. In Fig. 3d (middle), T^* depends quadratically on V_{BWT} away from the QCP, although the curvature differs between the two sides of the QCP, which have different ground states. The phase shift $\delta_P \approx 0$ on one side of the QCP, and appears to approach $\pi/2$ on the other (Fig. 3d, bottom). The $\pi/2$ shift reflects a sign flip in $S_{\sigma\alpha, \sigma\alpha}$ between distinct FL ground states, where either the grain or leads screen the dot spin²⁷. Between QCPs, δ_P varies smoothly. T^* and δ_P are not plotted directly to the right of each QCP, reflecting the ambiguity of fitting a small crossover peak on top of the 2CK peak. Both T^* and δ_P are insensitive to small changes in the range of V_{SD} used for fitting (Methods section ‘Sensitivity of T^* and δ_P to fitting range’).

A remarkable conclusion of these measurements is that, contrary to prevailing belief, the neighbourhood of a 2CK QCP can be quite large.

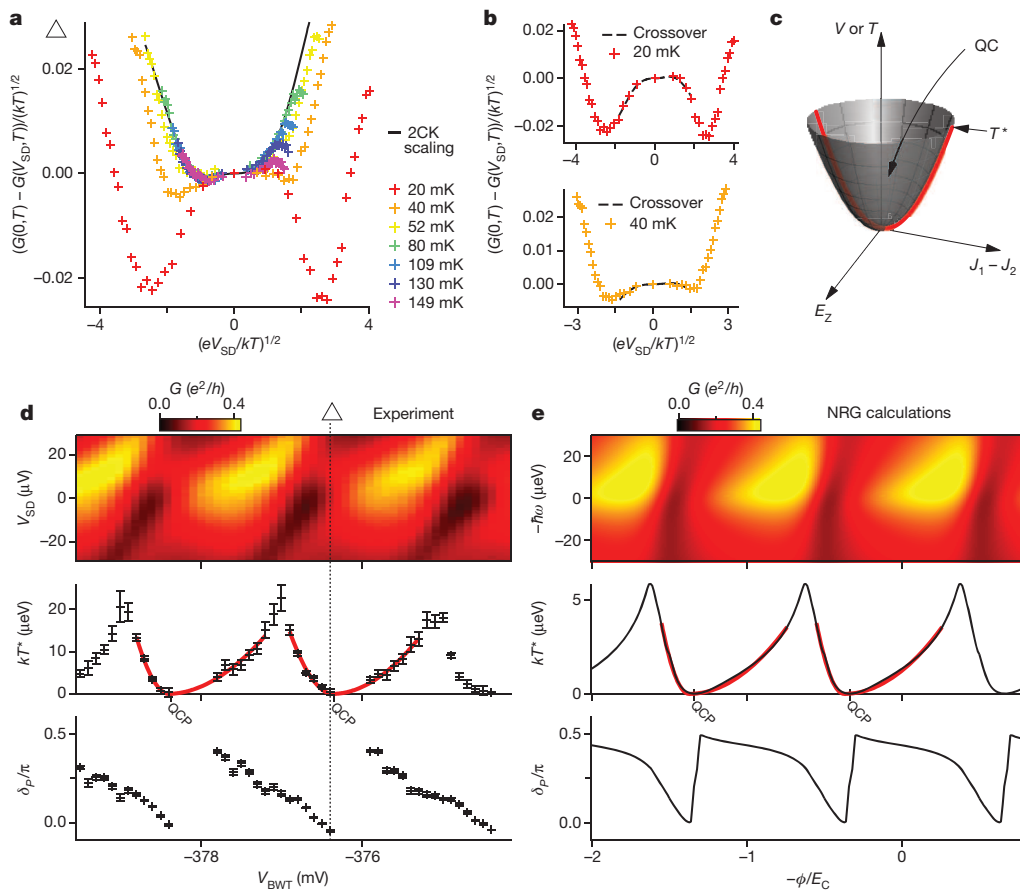


Figure 3 | Crossover from quantum criticality to a Fermi liquid.

$V_{LP} = -236.0$ mV for experimental data. **a**, Measured $G(V_{SD}, T)$. At $V_{BWT} = -376.4$ mV, a thermally broadened spectral function from the 2CK CFT ($\delta_P = -0.022\pi$, solid black line) describes the high- T data. **b**, Subset of data shown in **a**, with fits. $G(V_{SD})$ at low energies is fitted to thermally broadened spectral functions from the crossover theory (top, 20 mK; bottom, 40 mK; $\delta_P = -0.045\pi$, $T^* = 0.5$ μ eV). Fitting details in Methods. **c**, Quantum criticality (QC) occurs for energies above the Fermi liquid scale T^* (grey paraboloid), which should depend quadratically on the coupling asymmetry $J_1 - J_2$ between the two channels as well as on the Zeeman splitting E_Z . We vary

T^* by tuning $J_1 - J_2$ (cut along red parabola). **d**, Extraction of T^* and δ_P from measurements. The triangle denotes V_{BWT} for **a** and **b**. Top, $G(V_{SD}, V_{BWT})$ ($T = 20$ mK). Middle, T^* from crossover theory fits to the experimental $G(V_{SD}, T)$. Red traces are parabolas with $T^* = 0$ at the QCP and unequal scale factors on either side of the QCP. The largest T^* values may not be much less than T_K , so the crossover theory is not strictly valid for all V_{BWT} . Labels indicate approximate QCP locations. Bottom, δ_P from the crossover theory fits. Error bars, 1 s.d. confidence intervals from the fits. **e**, Extraction of T^* and δ_P from NRG calculations. Parameters as in Fig. 2. Top, $G(-\hbar\omega, -\phi/E_C)$, rescaled to match maximum G of **d**. Middle, T^* . Bottom, δ_P .

The T^* parabolas span most of the range in V_{BWT} , so fine tuning is not needed to reach a state within the influence of a QCP. Also, since $T^* \propto |J_1 - J_2|^2$, the measurements imply the exchange couplings depend linearly on gate voltage, which is not usually anticipated. Prior theoretical studies on the model dot-grain Hamiltonian have not noted either of these features. Fitting the crossover theory to spectral functions from NRG yields conductance via equation (3) (Methods section ‘Relationship of $G(V_{\text{SD}}, T)$ to spectral functions’). The NRG conductance (Fig. 3e, top) shows zero bias dips transitioning into peaks, as well as the shift of the peak towards positive $-\hbar\omega$, as in transport spectroscopy (Fig. 3d). Importantly, the ϕ dependence of T^* (Fig. 3e, middle) shows extended parabolas like in the measurements, and the extracted δ_P (Fig. 3e, bottom) reproduces the rapid $\pi/2$ phase shift across a QCP, with an otherwise smooth ϕ dependence. A perfect correspondence between experiment and NRG calculations should not be expected, since not all parameters may be extracted directly from measurements, and the dot-grain Hamiltonian is an idealization (Methods sections ‘Hamiltonian’ and ‘Extracting device parameters’). Yet the qualitative numerical reproduction of key experimental features helps to validate our surprising conclusions.

The experimental and numerical corroboration of analytical results in the vicinity of a QCP is a milestone in our understanding of correlated electron systems, with implications for other systems that may be influenced by a QCP, such as high- T_c superconductors and heavy fermion materials. An essentially identical universal crossover (same $\mathcal{G}(\hbar\omega/kT^*, T/T^*)$, but different symmetry-breaking perturbations) is expected for the two-impurity Kondo model. Future work will address the full phase diagram of the device, which may host charge 2CK^{23,25,28} and SU(4) Kondo regimes^{29,30}. Our device geometry could enable Aharonov–Bohm interference measurements to probe phase coherence of low-lying excitations in the non-FL 2CK state^{27,31}, giving new insight into the nature of a local non-FL.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 11 March; accepted 28 July 2015.

- Sachdev, S. *Quantum Phase Transitions* 2nd edn (Cambridge Univ. Press, 2011).
- Oreg, Y. & Goldhaber-Gordon, D. Two-channel Kondo effect in a modified single electron transistor. *Phys. Rev. Lett.* **90**, 136602 (2003).
- Potok, R. M., Rau, I. G., Shtrikman, H., Oreg, Y. & Goldhaber-Gordon, D. Observation of the two-channel Kondo effect. *Nature* **446**, 167–171 (2007).
- Nozières, Ph. & Blandin, A. Kondo effect in real metals. *J. Phys. (Paris)* **41**, 193–211 (1980).
- Sela, E., Mitchell, A. K. & Fritz, L. Exact crossover Green function in the two-channel and two-impurity Kondo models. *Phys. Rev. Lett.* **106**, 147202 (2011).
- Mitchell, A. K. & Sela, E. Universal low-temperature crossover in two-channel Kondo models. *Phys. Rev. B* **85**, 235127 (2012).
- Gegenwart, P., Si, Q. & Steglich, F. Quantum criticality in heavy-fermion metals. *Nature Phys.* **4**, 186–197 (2008).
- Coleman, P., Pépin, C., Si, Q. & Ramazashvili, R. How do Fermi liquids get heavy and die? *J. Phys. Condens. Matter* **13**, R723–R738 (2001).
- Mebrahtu, H. T. et al. Quantum phase transition in a resonant level coupled to interacting leads. *Nature* **488**, 61–64 (2012).
- Mebrahtu, H. T. et al. Observation of Majorana quantum critical behaviour in a resonant level coupled to a dissipative environment. *Nature Phys.* **9**, 732–737 (2013).
- Jeong, H., Chang, A. M. & Melloch, M. R. The Kondo effect in an artificial quantum dot molecule. *Science* **293**, 2221–2223 (2001).
- Bork, J. et al. A tunable two-impurity Kondo system in an atomic point contact. *Nature Phys.* **7**, 901–906 (2011).
- Chorley, S. J. et al. Tunable Kondo physics in a carbon nanotube double quantum dot. *Phys. Rev. Lett.* **109**, 156804 (2012).
- Emery, V. J. & Kivelson, S. Mapping of the two-channel Kondo problem to a resonant-level model. *Phys. Rev. B* **46**, 10812–10817 (1992).
- Cox, D. L. Quadrupolar Kondo effect in uranium heavy-electron materials? *Phys. Rev. Lett.* **59**, 1240–1243 (1987).
- Seaman, C. L. et al. Evidence for non-Fermi-liquid behavior in the Kondo alloy $\text{Y}_{1-x}\text{U}_x\text{Pd}_3$. *Phys. Rev. Lett.* **67**, 2882–2885 (1991).
- Besnus, M. J. et al. Specific heat and NMR of the Kondo system YbPd_2Si_2 . *J. Magn. Magn. Mater.* **76–77**, 471–472 (1988).
- Ralph, D. C., Ludwig, A. W. W., von Delft, J. & Burhman, R. A. 2-channel Kondo scaling in conductance signals from 2-level tunneling systems. *Phys. Rev. Lett.* **72**, 1064–1067 (1994).
- Cichorek, T. et al. Two-channel Kondo effect in glasslike ThAsSe . *Phys. Rev. Lett.* **94**, 236603 (2005).
- Yeh, S.-S. & Lin, J.-J. Two-channel Kondo effects in $\text{Al}/\text{AlO}_x/\text{Sc}$ planar tunnel junctions. *Phys. Rev. B* **79**, 012411 (2009).
- Tóth, A. I., Borda, L., von Delft, J. & Zaránd, G. Dynamical conductance in the two-channel Kondo regime of a double dot system. *Phys. Rev. B* **76**, 155318 (2007).
- Fabrizio, M., Gogolin, A. O. & Nozières, Ph. Anderson-Yuval approach to the multichannel Kondo problem. *Phys. Rev. B* **51**, 16088–16097 (1995).
- Anders, F. B., Lebanon, E. & Schiller, A. Coulomb blockade and non-Fermi-liquid behavior in quantum dots. *Phys. Rev. B* **70**, 201306(R) (2004).
- Anders, F. B., Lebanon, E. & Schiller, A. Conductance in coupled quantum dots: indicator for a local quantum phase transition. In *NIC Symposium Vol. 32*, 191–199 (John von Neumann Institute for Computing, Jülich, 2006).
- Lebanon, E., Schiller, A. & Anders, F. B. Enhancement of the two-channel Kondo effect in single-electron boxes. *Phys. Rev. B* **68**, 155301 (2003).
- Affleck, I. & Ludwig, A. W. W. Exact conformal-field-theory results on the multichannel Kondo effect: single-fermion Green’s function, self-energy, and resistivity. *Phys. Rev. B* **48**, 7297–7321 (1993).
- Borda, L., Fritz, L., Andrei, N. & Zaránd, G. Theory of inelastic scattering from quantum impurities. *Phys. Rev. B* **75**, 235112 (2007).
- Matveev, K. A. Quantum fluctuations of the charge of a metal particle under the Coulomb blockade conditions. *Sov. Phys. JETP* **72**, 892–899 (1991).
- Le Hur, K., Simon, P. & Borda, L. Maximized orbital and spin Kondo effects in a single-electron transistor. *Phys. Rev. B* **69**, 045326 (2004).
- Le Hur, K., Simon, P. & Loss, D. Transport through a quantum dot with SU(4) Kondo entanglement. *Phys. Rev. B* **75**, 035332 (2007).
- Carmi, A., Oreg, Y., Berkooz, M. & Goldhaber-Gordon, D. Transmission phase shifts of Kondo impurities. *Phys. Rev. B* **86**, 115129 (2012).

Acknowledgements We are grateful to S. Amasha, Y. Oreg, A. Carmi, E. Sela, A. K. Mitchell and M. Heiblum for discussions; H. K. Choi, Y. Chung and J. MacArthur for electronics expertise; M. Heiblum for use of his laboratory during initial device characterization; H. Inoue, N. Ofek, O. Raslin and E. Weisz for fabrication guidance; F. B. Anders, E. Lebanon and the late A. Schiller for their calculations which guided prior experimental work; and M. Stopa for his SETe software for electrostatic quantum dot modelling. The device was fabricated in the Braun Submicron Center at the Weizmann Institute of Science, with final fabrication steps done at Stanford Nano Shared Facilities (SNSF) at Stanford University. This work was supported by the Gordon and Betty Moore Foundation grant no. GBMF3429, the Hungarian research grant OTKA K105149, the Polish National Science Centre project no. DEC-2013/10/E/ST3/00213, EU grant no. CIG-303 689, the National Science Foundation grant no. DMR-0906062, and the US-Israel BSF grant no. 2008149. A.J.K. and L.P. were supported by a Stanford Graduate Fellowship. SETe calculations were run on the Odyssey cluster supported by the FAS Division of Science, Research Computing Group at Harvard University. NRG calculations were performed at the Poznań Supercomputing and Networking Center.

Author Contributions A.J.K., G.Z. and D.G.-G. designed the experiment. A.J.K. and L.P. performed the measurements. I.W., C.P.M. and G.Z. performed the NRG calculations. C.P.M. and I.W. contributed equally to the theoretical analysis. A.J.K., L.P., C.P.M., I.W., G.Z. and D.G.-G. analysed the data. A.J.K. designed and fabricated the devices, with e-beam lithography from D.M., using heterostructures grown by V.U. A.J.K. and L.P. wrote the paper with critical review provided by all other authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.G.-G. (goldhaber-gordon@stanford.edu).

METHODS

Device. The 2DEG is 50 nm below the surface and has an electron density $n = 3.3 \times 10^{11} \text{ cm}^{-2}$ and mobility $\mu = 1.2 \times 10^6 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$.

Figure 1a shows a top-down SEM micrograph of the device. This view is appropriate for labelling the gate electrodes and explaining the function of each gate, but the air bridges are hard to see. In Extended Data Fig. 1 we show a view of the device at a 40° tilt with respect to normal incidence. The five air bridges clearly rise above the gate electrodes underneath. The device is rotated approximately 180° with respect to the orientation of Fig. 1a.

As initially fabricated, the bridges did not make good electrical contact to the gates. This problem was remedied with an *in situ* platinum deposition procedure³². The SEM images of Fig. 1a and Extended Data Fig. 1 were obtained before the platinum deposition.

Measurements. The measurements are performed in the mixing chamber of a wet dilution refrigerator (Oxford Kelvinox TLM) with a base electron temperature $T_e = 20 \text{ mK}$, verified by Coulomb blockade thermometry. The device was cooled down with $+300 \text{ mV}$ bias on all gates to enhance charge stability by reducing the range of voltage needed for operation³³. Remaining charge instability manifested in discrete jumps of features in gate voltage every few hours, which were compensated by simply relocating those features with other parameters (for example, temperature) fixed. For all measurements we use an SR830 lock-in amplifier with $1 \mu\text{V}$ excitation at 33 Hz and a custom 10^8 V A^{-1} gain current preamplifier (design by H. K. Choi and Y. Chung, see related publication³⁴). A custom voltage source with six 20-bit channels and eight 16-bit channels was used to control the gate and source–drain bias voltages (design by J. MacArthur, assembled and calibrated by A.J.K.).

The biased source lead in any source–drain bias spectroscopy was determined to be weakly coupled to the dot: at zero bias, we pinch off the source lead’s coupling to the dot Γ_s (for example, using V_{LWT}) and observe a decrease in the overall conductance scale, without appreciable changes in the conductance features after accounting for capacitive shifts from gating. This implies that the unbiased drain lead’s coupling to the dot Γ_D largely determines the total dot–lead coupling rate, since $\Gamma = \Gamma_s + \Gamma_D \approx \Gamma_D$, that is, the dot was nearly in equilibrium. In comparing Fig. 2a and c, the ratio of maximum conductances is 0.464. If these numerical and experimental data areas are assumed to be directly comparable, then the asymmetry prefactor $4\Gamma_s\Gamma_D/(\Gamma_s + \Gamma_D)^2 = 0.464$, yielding $\Gamma_s/\Gamma_D = 15\%$. We consider this ratio small enough to compare dI/dV_{SD} with equilibrium spectral functions, as in the following section. As justification, consider that the conductance in Fig. 3e is derived from NRG calculations of equilibrium spectral functions as described in the following section, and reflect what would be anticipated in a measurement with the source lead coupled extremely weakly. Fitting the crossover theory to these spectral functions yields a dependence of T^* on ϕ that is very similar to that observed in experiments, up to scale factors which would be hard to attribute to any one cause.

It is well known that applying source–drain bias will cause unintentional gating as a secondary effect. This would be deleterious to observing quantum critical behaviour, which depends sensitively on the dot and grain levels. We compensate for shifts in the grain level by compensating changes in V_{SD} with changes in V_{BWT} . This compensation can be determined easily in the regime $e/U > 0$ or $e/U < -1$. We expect the grain level to be much more sensitive than the dot level for the same change in energy since $E_c \ll U$.

Relationship of $G(V_{\text{SD}}, T)$ to spectral functions. The differential conductance $G = dI/dV_{\text{SD}}$ measured from source to drain lead through the small dot is a function of source–drain bias and temperature and can be compared directly to NRG calculations in the case of a strongly asymmetric source–drain coupling. In the case of weak coupling to the biased source electrode ($\Gamma_s \ll \Gamma_D$), the differential conductance can be related to the equilibrium spectral function as

$$G(V_{\text{SD}}, T) \approx \frac{2e^2}{h} \frac{4\Gamma_s\Gamma_D}{(\Gamma_s + \Gamma_D)^2} \int_{-\infty}^{\infty} d\omega \left(-\frac{\partial f(\omega - (-eV_{\text{SD}}/\hbar), kT/\hbar)}{\partial \omega} \right) A(\omega, T) \quad (3)$$

The asymmetry prefactor is a function of the source and drain couplings, Γ_s and Γ_D , and is assumed to be much smaller than one. Either lead may assume the role of source or drain. The derivative of the Fermi–Dirac distribution $f(\omega, T)$ is convolved with a spectral function $A(\omega, T)$ from the 2CK or crossover descriptions. The spectral function can be related to the \mathcal{T} matrix

$$A(\omega, T) = -\pi v \sum_{\sigma} \text{Im}[\mathcal{T}_{\sigma\alpha, \sigma\alpha}(\omega, T)]|_{\alpha=-1} \quad (4)$$

where v is the bare density of states in the leads, σ is a spin index, and α is a channel index (we fix $\alpha = -1$ for the source and drain leads). The \mathcal{T} matrix represents the scattering between different states induced by the interaction part of the

Hamiltonian and can be computed numerically exactly by NRG. It is related to the quasiparticle self-energy³⁵.

Fitting expressions for 2CK. In equilibrium, the conduction electrons’ scattering \mathcal{T} matrix is proportional to the self-energy. In case of the quantum dot system considered here, the latter quantity translates to the Green’s function of the d -level of the small dot. This allows us to use the exact S matrix at the 2CK fixed point²⁶ and express the equilibrium spectral function of the small dot in the limit $kT^* \ll \hbar\omega$, $kT \ll kT_K$ as

$$A(\omega, T) \approx A_{2\text{CK}}(\omega, T, \delta_p) = \text{Im} i \left(1 - 3\lambda e^{2i\delta_p} \sqrt{\frac{\pi T}{T_K}} \int_0^1 du \left\{ u^{-i\beta\hbar\omega/2\pi} (1-u)^{1/2} \times u^{-1/2} {}_2F_1(3/2, 3/2; 1, u) - \frac{4}{\pi} u^{-1/2} (1-u)^{-3/2} \right\} \right) \quad (5)$$

where ${}_2F_1(a, b; c, z)$ is the Gauss hypergeometric function, β is $(kT)^{-1}$ and δ_p is the scattering phase shift. We fix the dimensionless parameter $\lambda = -0.09$ so that the spectral function drops to half of its $\omega = 0$ value at $\hbar\omega = kT_K$ in the limit $T \rightarrow 0$ (refs 21, 36). Equation (5) immediately implies that $(A_{2\text{CK}}(0, T, \delta_p) - A_{2\text{CK}}(\omega, T, \delta_p))/\sqrt{T_K/T}$ is a universal function of $\hbar\omega/kT$, which when convolved with a Fermi function gives the function $Y_{2\text{CK}}(-eV_{\text{SD}}/kT, \delta_p)$ of equation (1). We stress that this $\hbar\omega/kT$ scaling is a special property of the 2CK fixed point. When fitting the experimental data, we shall assume an asymmetric coupling to the leads (see previous section).

Fitting expressions for crossover. At frequencies and temperatures far below the 2CK temperature T_K , we can use the crossover form of the \mathcal{T} matrix derived in refs 5 and 6 to express the d -level’s equilibrium spectral function. Here we obtain the expression

$$A(\omega, T) \approx A_{\text{FL}}(\omega, T, \delta_p) \equiv \text{Im} i (1 - e^{2i\delta_p} \mathcal{G}(\tilde{\omega}, \tilde{T})) \quad (6)$$

where $\delta_p \approx 0$ ($\delta_p \approx \pi/2$) in the case when the dot is coupled more strongly to the grain (leads), and

$$\mathcal{G}(\tilde{\omega}, \tilde{T}) = \frac{-i}{\sqrt{2\pi^3 \tilde{T}}} \Gamma\left(\frac{1}{2} + \frac{1}{2\pi\tilde{T}}\right) \frac{\tanh \frac{\tilde{\omega}}{2\tilde{T}}}{\Gamma\left(1 + \frac{1}{2\pi\tilde{T}}\right)} \times \int_{-\infty}^{+\infty} dx \frac{e^{ix\tilde{\omega}/\pi\tilde{T}}}{\sinh x} \text{Re} \left[{}_2F_1\left(\frac{1}{2}, \frac{1}{2}; 1 + \frac{1}{2\pi\tilde{T}}, \frac{1 - \coth x}{2}\right) \right] \quad (7)$$

is a universal function of rescaled energy $\tilde{\omega} = \hbar\omega/kT^*$ and temperature $\tilde{T} = T/T^*$. For equation (7) only, Γ is the gamma function, not a tunnel rate. Again, when fitting to experimental data, the spectral function must be thermally broadened (see Methods section ‘Relationship of $G(V_{\text{SD}}, T)$ to spectral functions’).

Fitting range. When fitting the crossover theory to experimental data, we fit $G(V_{\text{SD}}, T)$ only in a small window of V_{SD} of $\pm 6 \mu\text{V}$ around zero, regardless of temperature. A priori, T^* is unknown and it only makes sense to fit $eV_{\text{SD}} < \text{a few } kT^*$. Additionally, thermal broadening of high energy features can in principle spoil the scaling of the low energy features, even for otherwise sensible ranges of V_{SD} . At minimum the 20 mK and 40 mK traces are used for fitting, but sometimes also the 52 mK and possibly the 70 mK traces, provided $T \lesssim T^*$ (the fitting process is somewhat iterative in this respect). Once the temperatures to be used in fitting are decided for a given value of V_{BWT} , the fitting considers data from all of those temperatures simultaneously. Fitting the crossover theory to NRG calculations is done analogously (a window of $\hbar\omega$ of $\pm 6 \mu\text{eV}$ about zero).

Sensitivity of T^* and δ_p to fitting range. In Fig. 3 we use the crossover CFT to fit experimental data and thereby extract the Fermi liquid scale T^* and the scattering phase shift δ_p . The fitting procedure uses a limited range of V_{SD} ($\pm 6 \mu\text{V}$) and it is argued that this is a conservative approach.

In Extended Data Fig. 2 we show that the fitting is insensitive to small changes in the fitting range. At each value of V_{BWT} , we try and extract T^* and δ_p for nine different ranges of bias voltage, which we obtain by starting with $(-6, +6 \mu\text{V})$ and adding or subtracting a point on either end, for example, $(-7.5, +6)$, $(-6, +6)$, $(-6, +7.5)$, $(-4.5, +6)$ and so on. It is important not to add so many points that data outside the validity of the theory are included. However, subtracting too many points may degrade the fit quality.

After finding T^* and the error in T^* reported by the fits, we consider all nine fitting ranges to give independent estimates of T^* , and find the weighted mean T^* , weighted by the errors from each fit. The error bars show the standard deviation of the weighted mean, and indicate the spread of T^* values returned by the fits. We do the same for δ_p . Varying the fitting range by small amounts does not seem to contribute significantly to the uncertainty in T^* and δ_p .

Scaling along the quantum critical lines. In Fig. 2e we demonstrated that measured $G(V_{SD}, T)$ falls onto a scaling curve derived from the CFT results of ref. 26. However, at the lowest temperatures there are deviations from the expected scaling curve. Here we justify attributing the deviations to a small $T^* \lesssim T_c$. For clarity we break out the data of Fig. 2e into separate panels for each temperature in Extended Data Fig. 3.

For $T \gtrsim 80$ mK, the data appear to collapse onto the scaling curve. Deviations appear for $V_{SD} > 0$ and become increasingly prominent as temperature decreases. Focusing on the 20 mK trace, the difference between the data and scaling curve vanishes by construction at $V_{SD} = 0$, but becomes quite large for small positive V_{SD} . The difference between the scaling curve and the data actually decreases at larger V_{SD} . These observations are at least qualitatively consistent with the effect of a small T^* , which should result in corrections at low energies. It is perhaps surprising that a small $T^* \lesssim T$ could result in substantial deviations from 2CK scaling, but we estimate from figure 7 of ref. 6 that even for $T \approx 10T^*$, signs of the crossover may be easily seen up to $eV_{SD} \approx 30kT^*$, and perhaps even factors of a few higher than that.

It is notable that the deviations from 2CK scaling appear significant only for $V_{SD} > 0$, which would seem to imply a scattering phase that is not quite $\pi/2$. Another possibility to explain the asymmetry about $V_{SD} = 0$ would be that true zero V_{SD} drifted over the course of the measurement. Typically some small applied V_{SD} is required to compensate for an offset voltage at the current amplifier input, but at base temperature the quality of the scaling collapse is sensitive to errors of the order of 1 μ V in identification of true zero V_{SD} .

In Extended Data Fig. 4 we demonstrate the 2CK scaling at other points along the 2CK lines. In most examples, the 2CK scaling behaviour is faithfully reproduced by the data except perhaps at $T = 20$ mK. This may be considered experimental evidence for the existence of the 2CK lines.

Electron temperature. We determine the electron temperature in our system using the small quantum dot as a Coulomb blockade thermometer, during the same cooldown of the device in which the data of Fig. 2 and Fig. 3 were measured. For a dot with level spacing ΔE , charging energy e^2/C and total tunnel rate Γ to its leads, the conductance lineshape at a temperature T is given by³⁷

$$G = G_0 \frac{e^2}{4kT} \cosh^{-2} \left(\frac{e}{2kT} \right) \quad \text{for} \quad \hbar\Gamma \ll kT \ll \Delta E, e^2/C \quad (8)$$

where e is the dot level and G_0 is an asymmetry prefactor that depends on the tunnel rates between the dot and each lead to which it is coupled. Changes in voltage on gate LP are related to changes in the dot level via the lever arm α_{LP} : $\Delta e = -\alpha_{LP}|e|\Delta V_{LP}$. The slopes m and n of the prominent diagonal features in source-drain bias spectroscopy (Extended Data Fig. 5a) are used to obtain $\alpha_{LP} = -mn/(m-n) = 0.0786$. Measured temperature-dependent conductance is fitted with equation (8) (Extended Data Fig. 5b), yielding the electron temperatures as stated in the legend. The base electron temperature $T_c = 20$ mK, and elevated electron temperatures are achieved by continuously heating a resistor in the mixing chamber of the dilution refrigerator and waiting for equilibration. By calibrating ruthenium oxide resistance thermometers near the device to the measured electron temperatures, the electron temperature can be inferred during the measurements of Fig. 2 and Fig. 3, in which the dot is not in the regime of strong Coulomb blockade.

In equation (8), the peak height is predicted to have a T^{-1} dependence. We plot the peak height versus the electron temperature in Extended Data Fig. 5c. A power law fit yields an exponent of $-1.04(1)$, reasonably close to the expected value given that less than an order of magnitude in temperature is considered. No obvious systematic error is seen in the residuals which are all less than $0.001e^2/h$ (Extended Data Fig. 5d).

Hamiltonian. In our numerical calculations the dot-grain system is modelled by the Hamiltonian

$$H_{\text{device}} = H_{\text{dot}} + H_{\text{grain}} + H_{\text{leads}} + H_{\text{tunnelling}} \quad (9)$$

where

$$H_{\text{dot}} = \sum_{\sigma} \varepsilon d_{\sigma}^{\dagger} d_{\sigma} + U \hat{n}_1 \hat{n}_1 \quad (10)$$

describes the dot, with ε the on-site energy, d_{σ}^{\dagger} the creation operator of an electron with spin σ and $\hat{n}_{\sigma} = d_{\sigma}^{\dagger} d_{\sigma}$ representing the occupation number operator. U is the correlation energy between the two electrons residing in the dot. The grain is described by

$$H_{\text{grain}} = \sum_{p,\sigma} \varepsilon_p a_{p\sigma}^{\dagger} a_{p\sigma} + \frac{E_C}{2} (\hat{n}_g - N_0)^2 + \phi (\hat{n}_g - N_0) \quad (11)$$

The creation operator of a spin- σ electron with momentum p and energy ε_p in the grain is denoted by $a_{p\sigma}^{\dagger}$, E_C is the charging energy of the grain, while $-\phi$ plays the

role of a gate voltage. \hat{n}_g is the electron number operator of the grain, $\hat{n}_g = \sum_{p,\sigma} a_{p\sigma}^{\dagger} a_{p\sigma}$, and N_0 denotes the number of excess electrons in the electrically neutral grain ($\phi = 0$). The non-interacting quasiparticles in the leads are described by

$$H_{\text{leads}} = \sum_{\alpha=\{S,D\}} \sum_{k,\sigma} \varepsilon_{\alpha k} c_{\alpha k\sigma}^{\dagger} c_{\alpha k\sigma} \quad (12)$$

where $c_{\alpha k\sigma}^{\dagger}$ is the creation operator of a spin- σ electron with momentum k and energy $\varepsilon_{\alpha k}$ in the source ($\alpha = S$) or drain ($\alpha = D$) lead. The last term in equation (9) is the tunnelling Hamiltonian, which is given by

$$H_{\text{tunnelling}} = t_G \sum_{p,\sigma} \left(a_{p\sigma}^{\dagger} d_{\sigma} + d_{\sigma}^{\dagger} a_{p\sigma} \right) + \sum_{\alpha=\{S,D\}} \sum_{k,\sigma} t_{\alpha} \left(c_{\alpha k\sigma}^{\dagger} d_{\sigma} + d_{\sigma}^{\dagger} c_{\alpha k\sigma} \right) \quad (13)$$

The tunnel matrix elements between the leads (grain) and the dot are denoted by t_{α} (t_G) and are assumed to be independent of momentum. The strengths of the couplings are given by $\Gamma_G = \pi v_G |t_G|^2$ and $\Gamma_{\alpha} = \pi v_{\alpha} |t_{\alpha}|^2$, respectively, where $v_{\alpha}(v_G)$ is the density of states for lead α (grain). In the NRG calculations the energy spectrum of the grain is assumed to be continuous and the densities of states for leads and grain are taken to be constant and equal: $v_{\alpha} = v_G = v = 1/(2D)$, with $D \equiv 1$ being the half bandwidth used as the energy unit in NRG calculations.

In the Hamiltonian (equation (9)) we have neglected the dot-grain capacitive coupling, which can give rise to a term of the form, $U_{dg}(\hat{n}_g - N_0)\hat{n}_d$, where $\hat{n}_d = \hat{n}_1 + \hat{n}_1$. An estimate for U_{dg} is extracted experimentally in the Methods section 'Extracting device parameters', and U_{dg} is thought to play no role for the purpose of the present analysis.

Several other Hamiltonians have been proposed to realize 2CK physics in quantum dot based systems^{38,39}.

NRG calculations. To solve the Hamiltonian (equation (9)) we use the NRG method^{40,41} (open access Budapest code is available at <http://www.phy.bme.hu/~dmnrg/>). First, we introduce the collective charge operators (bosonic operators) for the grain^{23,25}

$$\hat{N} = \sum_{m=-\infty}^{\infty} m |m\rangle \langle m| \quad \text{and} \quad \hat{N}^{\pm} = \sum_{m=-\infty}^{\infty} |m \pm 1\rangle \langle m| \quad (14)$$

Strictly speaking, the identity, $\hat{N} = \hat{n}_g$ must be fulfilled, but within the NRG approach this constraint can be relaxed by treating \hat{N} as an independent quantity. This is possible as the spectral properties of the system are not sensitive to the exact number of conduction electrons present in the grain in the limit of infinitely small level spacing. To extract the finite size spectrum and determine the location of the 2CK lines, however, a projection to the physical subspace was necessary. In our calculations we took into account seven charges in the grain. Using the above charge operators, the grain part of the Hamiltonian (11) can be rewritten as

$$H_{\text{grain}} = \sum_{p,\sigma} \varepsilon_p a_{p\sigma}^{\dagger} a_{p\sigma} + \frac{E_C}{2} (\hat{N} - N_0)^2 + \phi (\hat{N} - N_0) \quad (15)$$

The \hat{N}^{\pm} operators capture the charging transitions of the grain and enter explicitly in the tunnelling Hamiltonian (equation (13)), which now reads

$$H_{\text{tunnelling}} = \sqrt{\frac{2\Gamma_G}{\pi}} \sum_{p,\sigma} \left(\hat{N}^{+} a_{p\sigma}^{\dagger} d_{\sigma} + d_{\sigma}^{\dagger} a_{p\sigma} \hat{N}^{-} \right) + \sqrt{\frac{2\Gamma}{\pi}} \sum_{k,\sigma} \left(c_{k\sigma}^{\dagger} d_{\sigma} + d_{\sigma}^{\dagger} c_{k\sigma} \right) \quad (16)$$

The first term in equation (16) describes the dot-grain tunnelling, while the second term accounts for the dot-leads coupling. This second term is obtained by performing an orthogonal transformation⁴² from the two-lead basis to an effective single lead with resultant coupling $\Gamma = \Gamma_S + \Gamma_D$. The resulting Hamiltonian consists then of two conduction bands coupled to a complex impurity composed of the grain (\hat{N}) and dot degrees of freedom.

The core of the NRG procedure is the logarithmic discretization of the conduction band with discretization parameter Λ and mapping of the conduction band onto a semi-infinite chain with exponentially decreasing hoppings. The Hamiltonian can then be diagonalized in an iterative fashion. In our calculations we used discretization parameter $\Lambda = 2$ and kept 4,000 states at each iteration. We also exploited the SU(2) symmetry of the total spin and two U(1) symmetries for $\hat{N}_1 = \hat{n}_d + \hat{n}_{cb} + \hat{n}_g$ and $\hat{N}_2 = \hat{n}_g - \hat{N}$, where \hat{n}_{cb} is the electron number operator in the first conduction channel (leads coupled to the dot) and \hat{n}_g is the electron number operator in the second channel (the grain). We performed the full density-matrix NRG calculations (fDM-NRG)^{40,43-45}, employing the Budapest Flexible DM-NRG code⁴¹, to compute the normalized dimensionless spectral function, $A(\omega, T) \equiv \pi(\Gamma_S + \Gamma_D)A_d(\omega, T)$, where $A_d(\omega, T)$ is the spectral

function for the d_{σ}^{\dagger} operators that describe the dot level. The linear conductance through the small dot can be then determined with the equation

$$G = \frac{2e^2}{h} \frac{4\Gamma_S\Gamma_D}{(\Gamma_S + \Gamma_D)^2} \int d\omega \left(-\frac{\partial f(\omega, kT/\hbar)}{\partial \omega} \right) A(\omega, T) \quad (17)$$

where $f(\omega, T)$ is the Fermi–Dirac distribution function.

Shifting of NRG calculations in Fig. 2b. In Fig. 2b we incorporate a linear ε -dependent shift into ϕ to obtain agreement between NRG calculations and experiment (Fig. 2c). The agreement is obtained by first rescaling the NRG calculations so that the maximum value of G is the same. The global scaling takes into account the source–drain coupling asymmetry, as explained elsewhere in Methods. Then, the sharp features in the cut taken at $V_{LP} = -260$ mV are compared with NRG calculations to establish $V_{LP} = -260$ mV $\approx -\varepsilon/U = 0.55$. Finally, another cut for fixed V_{LP} is taken to establish a linear relationship between V_{LP} and $-\varepsilon/U$. The two points give a linear relationship between V_{LP} and $-\varepsilon/U$. One global offset in $-\phi/E_C$ then suffices to give good agreement everywhere. Using this method we find $-\Phi/E_C = -\phi/E_C - 3.1(-\varepsilon/U - 1.5)$, where $-\Phi/E_C$ is the vertical axis of Fig. 2b. Physically, the linear dependence of Φ on $-\varepsilon/U$ can be understood as a consequence of the indirect capacitive coupling between V_{LP} and the grain.

Extracting device parameters. In this section we describe how model parameters are extracted from measurements, and comment on which of the parameters should be considered free. We also determine bounds on any dot–grain charging energy U_{dg} neglected in the model.

The dot charging energy $U = 2.9$ meV is determined from source–drain bias spectroscopy of the dot in the Coulomb blockade regime (Extended Data Fig. 6). In previous cooldowns U has varied between 1 and 3 meV, perhaps owing to how U depends sensitively on the number of electrons in the few electron regime. We use $U = 2$ meV as the model parameter in NRG calculations, and note that the calculations should be relatively insensitive when $U > D = 1$ meV, the electronic half bandwidth used in calculations. This value of D corresponds roughly to the internal level spacing on the small dot, providing a high energy cut-off.

The grain charging energy $E_C = e^2/C = 160$ μ eV is measured by source–drain bias spectroscopy of the grain (Extended Data Fig. 7). We compare this measurement to geometric estimates. A common rule of thumb is that upon gate depletion, the extent of the depletion region extends as far from the gate laterally as the 2DEG is deep. This means that the area of the grain should be the area outlined by the gates, less some area in an ~ 50 nm dead region adjacent to the gates. The red shaded area in Extended Data Fig. 7a is 1.2 μ m². Approximating the capacitance of the grain as that of a flat disk with radius r , $C = 8\epsilon r$ where $\epsilon = 13\epsilon_0$ for GaAs and the effective $r = 0.62$ μ m. This gives an expected $E_C = 280$ μ eV, which is within a factor of two of the measurement. In a previous cooldown of the same device we measured $E_C = 150$ μ eV, which we use as the model parameter in NRG calculations.

In designing the device we aimed for as large an E_C as possible while still being able to imagine a near continuum of states in the grain. The level spacing may be estimated by considering a particle in a 2D box. The level spacing $\Delta = \hbar^2\pi^2/2mA$, where A is the area of the box and $m = 0.067m_e$ is the effective mass in GaAs. Using the design area $A = 1.2$ μ m² we find $\Delta = 4.6$ μ eV $= 2.6kT_e$, where $T_e = 20$ mK. If we instead take $A = 0.93$ μ m² inferred from measurement of E_C and the approximation for the capacitance, we find $\Delta = 6.0$ μ eV $= 3.4kT_e$. In either case, Δ is no more than factors of a few times T_e , keeping in mind that the width of the Fermi–Dirac distribution is approximately $3.5kT_e$. This implies that the grain is indeed acting as a metallic grain at all measured temperatures. In Extended Data Fig. 7, it appears that the typical level spacing (spacing in V_{SD} between diagonal lines) is larger than anticipated. The peak conductance can differ significantly for each level, which reflects a distribution in source–drain coupling asymmetry from level to level. Some levels may not be visible if their source–drain coupling asymmetry is strong.

The dot level ε may be tuned by changing the voltage applied to gate LP. We typically tune V_{LP} by tens of mV, and only a change of 1.9 mV in V_{BWT} is needed to add an electron to the grain, so even a weak capacitance of gate LP to the grain may be important. Over a small range in V_{BWT} , the converse effect of V_{BWT} on the dot can be safely ignored. We think of ϕ as a function of both V_{LP} and V_{BWT} , and ε as a function of V_{LP} .

The dependence of ϕ on V_{BWT} may be determined easily. By tuning ε/U towards -1 using V_{LP} , the increasingly sharp features as a function of V_{BWT} (Fig. 2b, c) should be spaced by E_C in units of energy. These sharp features are rather generic as a function of all the other parameters. We already know E_C from a direct measurement, and therefore we have a conversion between V_{BWT} and ϕ . The dependence of ϕ on V_{LP} can be estimated by looking at the overall skew of features in Fig. 2c, and estimating some $\Delta V_{BWT}/\Delta V_{LP}$ to describe the skew. Estimating the skew is not strictly a well defined procedure, but the skew itself is well defined as the ratio of capacitances between gate LP and the grain and gate BWT and the grain. Since the dependence of ϕ on V_{BWT} is known explicitly and there is a good strategy for estimating the dependence on V_{LP} , we do not consider ϕ a free parameter.

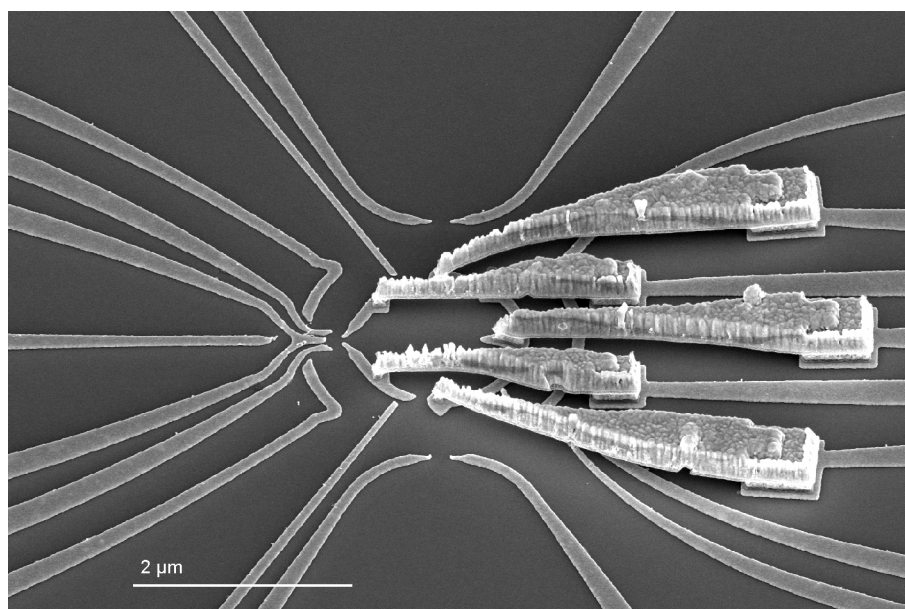
The dependence of ε on V_{LP} is hard to determine explicitly from measurements, although they should be proportional. Unlike how we calibrated V_{BWT} , there are no consistent sharp features as a function of V_{LP} which are independent of the other parameters. We believe we can identify to within the order of ten per cent the constant of proportionality just using the broad features observed in both experiments and calculations.

The tunnel couplings t_S , t_D and t_G are essentially free parameters in that we have not identified a way to extract them experimentally. Only an even combination of the source and drain leads will couple to the dot, so it makes more sense to think about the tunnel coupling for this even combination, t , and a source–drain asymmetry parameter, α , rather than t_S and t_D . Source–drain coupling asymmetry, in the limit of zero source–drain bias, just acts like a global scale factor. Apart from our intuition developed from experimental measurements, we estimate our source–drain coupling asymmetry by how much we need to scale the NRG calculations to match the experiment.

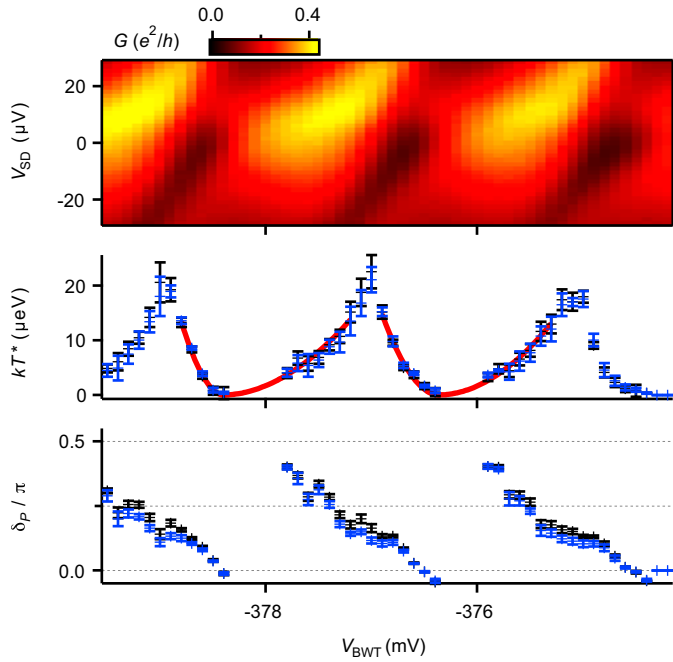
In summary, it would seem there are three free parameters: t , t_G , α . The last free parameter is a trivial scale factor. The other parameters can either be measured explicitly, or are linearly related to gate voltages and may be estimated well. We have tried exploring the (t, t_G) plane and cannot obtain much better agreement.

Finally, measurements of $G(V_{BWT}, V_{LP})$ yield $U_{dg} = 21$ μ eV (Extended Data Fig. 8). This analysis considers the dot–grain system as a capacitively-coupled double quantum dot. The U_{dg} we extract should be thought of as an upper bound—the gate voltages are set to a very different regime where Γ_G is negligible, unlike the situation in the work we report here. When tuning between this regime and the regime where $\Gamma_G \approx \Gamma$, it appears as if the splitting of the lines in Extended Data Fig. 8 goes to zero long before Γ_G becomes a significant fraction of Γ , perhaps implying that $U_{dg} \rightarrow 0$.

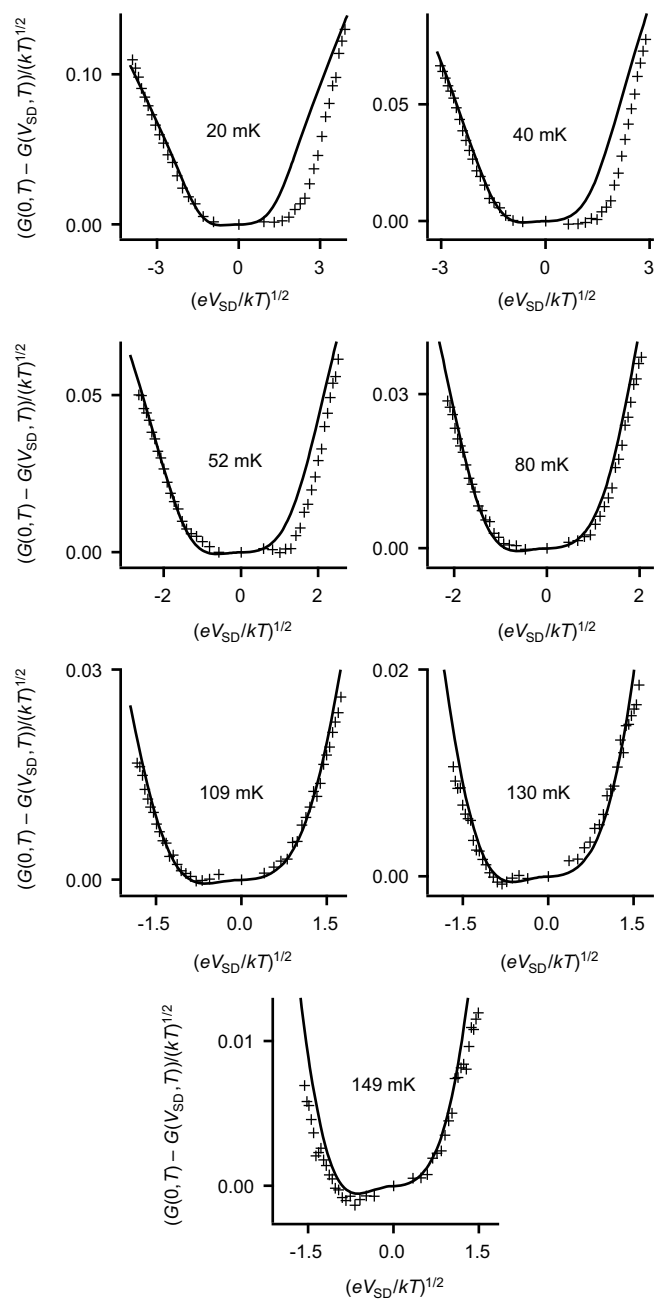
32. Peeters, L., Keller, A. J., Umansky, V., Mahalu, D. & Goldhaber-Gordon, D. Repairing nanoscale devices using electron-beam-induced deposition of platinum. *J. Vac. Sci. Technol. B* **33**, 051803 (2015).
33. Pioro-Ladrière, M. *et al.* Origin of switching noise in GaAs/Al_xGa_{1-x}As lateral gated devices. *Phys. Rev. B* **72**, 115331 (2005).
34. Kretinin, A. V. & Chung, Y. Wide-band current preamplifier for conductance measurements with large input capacitance. *Rev. Sci. Instrum.* **83**, 084704 (2012).
35. Altland, A. & Simons, B. D. *Condensed Matter Field Theory* 2nd edn (Cambridge Univ. Press, 2010).
36. Moca, C. P., Alex, A., von Delft, J. & Zaránd, G. SU(3) Anderson impurity model: a numerical renormalization group approach exploiting non-Abelian symmetries. *Phys. Rev. B* **86**, 195128 (2012).
37. Beenakker, C. W. J. Theory of Coulomb-blockade oscillations in the conductance of a quantum dot. *Phys. Rev. B* **44**, 1646–1656 (1991).
38. Bolech, C. J. & Shah, N. Prediction of the capacitance line shape in two-channel quantum dots. *Phys. Rev. Lett.* **95**, 036801 (2005).
39. Mitchell, A. K., Logan, D. E. & Krishnamurthy, H. R. Two-channel Kondo physics in odd impurity chains. *Phys. Rev. B* **84**, 035119 (2011).
40. Wilson, K. G. The renormalization group: critical phenomena and the Kondo problem. *Rev. Mod. Phys.* **47**, 773–840 (1975).
41. Legeza, Ö., Moca, C., Tóth, A., Weymann, I. & Zaránd, G. Manual for the Flexible DM-NRG Code. <http://arXiv.org/abs/0809.3143v1> (2008).
42. Pustilnik, M. & Glazman, L. I. Kondo effect in real quantum dots. *Phys. Rev. Lett.* **87**, 216601 (2001).
43. Bulla, R., Costi, T. A. & Pruschke, T. Numerical renormalization group method for quantum impurity systems. *Rev. Mod. Phys.* **80**, 395–450 (2008).
44. Weichselbaum, A. & von Delft, J. Sum-rule conserving spectral functions from the numerical renormalization group. *Phys. Rev. Lett.* **99**, 076402 (2007).
45. Tóth, A. I., Moca, C. P., Legeza, Ö. & Zaránd, G. Density matrix numerical renormalization group for non-Abelian symmetries. *Phys. Rev. B* **78**, 245109 (2008).



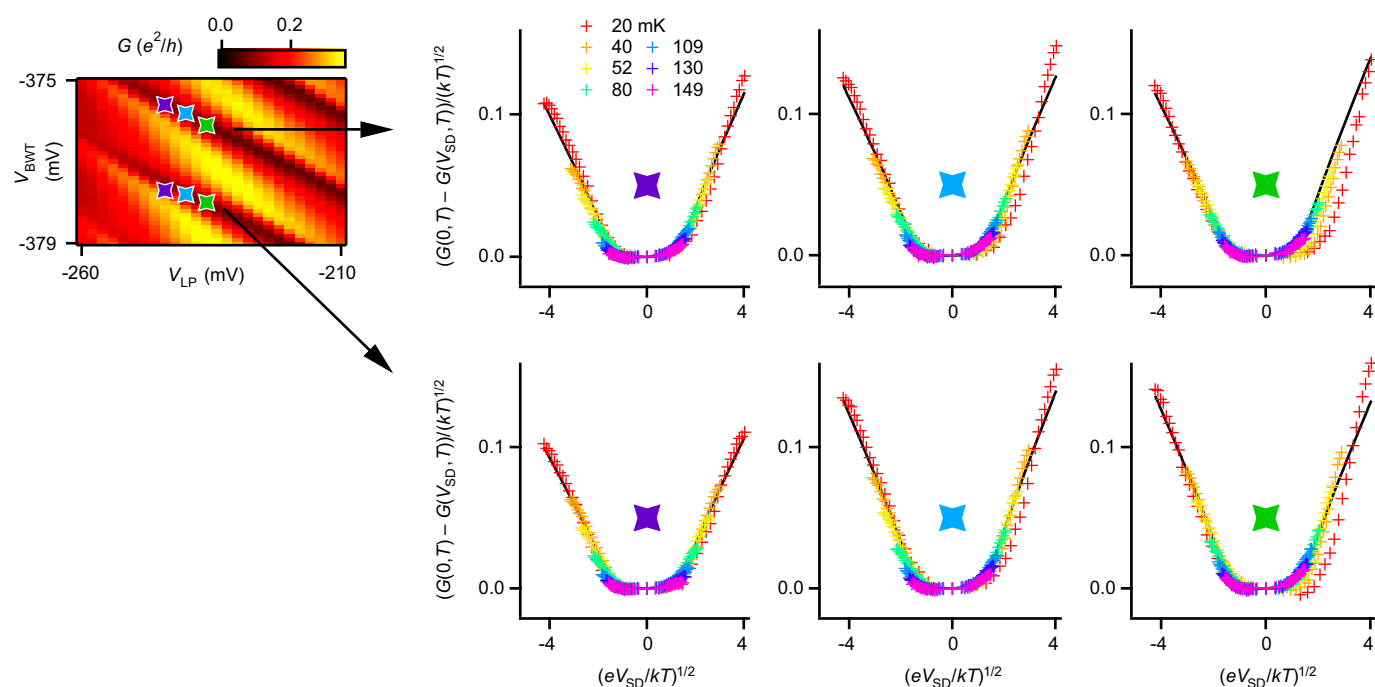
Extended Data Figure 1 | SEM micrograph of a device nominally identical to the device studied. Acceleration voltage in the SEM was 5 kV. The device is tilted 40° with respect to normal incidence.



Extended Data Figure 2 | Sensitivity of T^* and δ_p to fitting range. Top panel, $G(V_{SD}, V_{BWT})$ at $T = 20$ mK, exactly as in Fig. 3. Middle and bottom panels, T^* (middle) and δ_p (bottom) as functions of V_{BWT} . Black points and red curves are exactly as in Fig. 3; the blue points correspond to the weighted mean of extracted T^* and δ_p for an ensemble of fitting ranges, and error bars on the blue points correspond to the s.d. of the weighted mean.

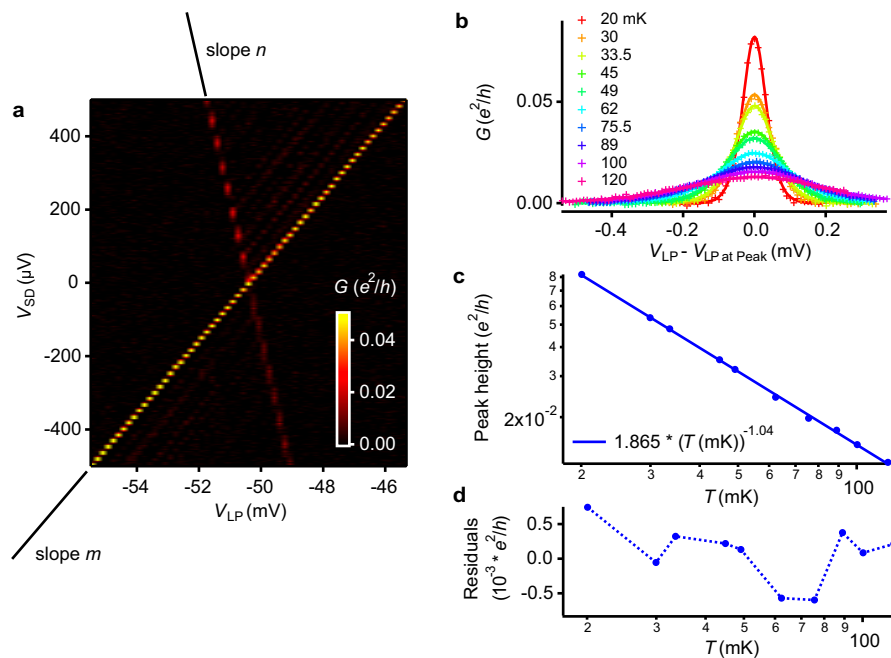


Extended Data Figure 3 | Measured $(G(0, T) - G(V_{SD}, T))/(kT)^{1/2}$ (symbols) and CFT fit (solid lines) of Fig. 2e broken out into separate panels for each T . Temperature T in mK is shown centrally in each panel. The range in measured V_{SD} is from -31.5 to 28.5 μV , resulting in a decreasing range on the $(eV_{SD}/kT)^{1/2}$ axis as T is increased. The single fit in Fig. 2e is plotted against the measured data for each T .



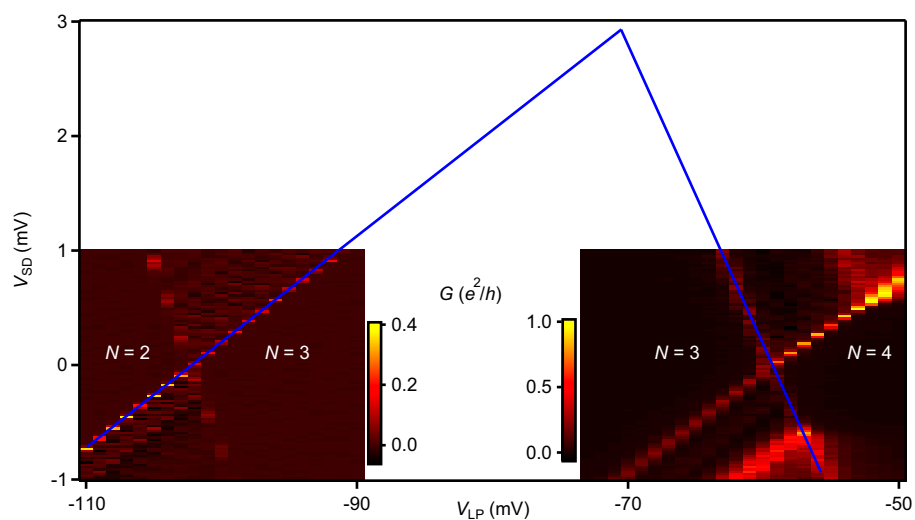
Extended Data Figure 4 | Two-channel Kondo scaling. Top left, measured $G(V_{LP}, V_{BWT})$ from Fig. 2c. Panels at right, measured $(G(0, T) - G(V_{SD}, T))/(kT)^{1/2}$ at six points on 2CK lines in the (V_{LP}, V_{BWT}) plane; points are indicated

by coloured stars. Black lines are fits to thermally broadened spectral functions from ref. 26 with small phase shifts from potential scattering.

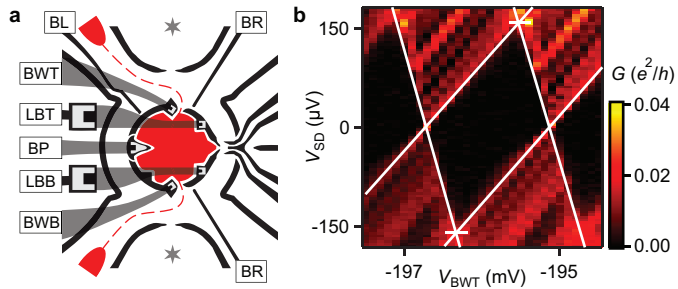


Extended Data Figure 5 | Coulomb blockade thermometry. **a**, Measured $G(V_{SD}, V_{LP})$ reveals two prominent linear features, the slopes of which are labelled as m and n . **b**, Measured $G(V_{SD} = 0, T)$ (crosses) and fits using equation (8) (lines). Every measurement is the average of 20 successive traces.

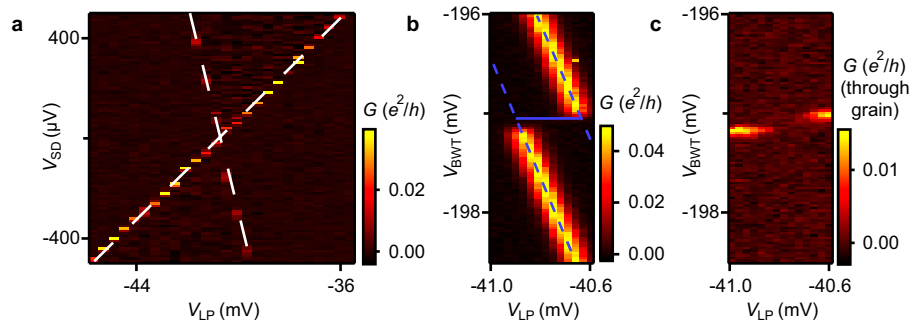
c, Power-law fit of peak height to extracted electron temperature yields an exponent of $-1.04(1)$. **d**, Residuals of the fit shown in **c** are all less than $0.001e^2/h$.



Extended Data Figure 6 | Measurement of U . Within each Coulomb diamond we label the number of electrons on the dot (N) as determined by charge sensing techniques. The intersection of the lines indicate $U \approx 2.9$ meV.



Extended Data Figure 7 | Measurement of E_C . **a**, Measurement scheme. $G(V_{SD}, V_{BWT})$ is measured using the grain's own pair of measurement leads (red pads), which are isolated from the measurement leads of the dot by depleting gate BR. Gate BL is depleted to avoid shorting conductance through the channel just left of the grain. The current path is the red dashed line. The grey stars indicate ohmic contacts which are floated during measurement. **b**, $G(V_{SD}, V_{BWT})$ through the grain in the Coulomb blockade regime ($T = 20$ mK). The intersections of the lines indicate $E_C \approx 160 \mu\text{eV}$.



Extended Data Figure 8 | Bounding U_{dg} from measurements in the Coulomb blockade regime. **a**, $G(V_{SD}, V_{LP})$ through the dot in the Coulomb blockade regime, with both the dot and grain formed. Here V_{BWT} is such that the grain is Coulomb-blockaded. From the slopes of the dashed lines overlaid on the peaks in the data, we determine lever arms $\alpha_{LP} = 0.081$ and $\alpha_{SD} = 0.194$. **b**, $G(V_{BWT}, V_{LP})$ through the dot at zero V_{SD} . Peaks in G correspond to Coulomb blockade on the dot being lifted; the splitting implies

finite U_{dg} . For fixed V_{BWT} the difference in peak positions (blue horizontal bar) gives the dot–grain charging energy $U_{dg} = (e)(\alpha_{LP})(\Delta V_{LP}) = 0.081 \times 0.26 \text{ meV} = 21 \text{ } \mu\text{eV}$. Dot–grain tunnelling is negligible in this limit. **c**, Conductance through the grain appears where expected given the interpretation of **b**. The conductance is measured with gates BL and BR depleted, measuring through the two point contacts formed by gate pairs LBB/BWB and LBT/BWT.

Identification of carbohydrate anomers using ion mobility–mass spectrometry

J. Hofmann^{1,2*}, H. S. Hahm^{3,4*}, P. H. Seeberger^{3,4} & K. Pagel^{1,2}

Carbohydrates are ubiquitous biological polymers that are important in a broad range of biological processes^{1–3}. However, owing to their branched structures and the presence of stereogenic centres at each glycosidic linkage between monomers, carbohydrates are harder to characterize than are peptides and oligonucleotides⁴. Methods such as nuclear magnetic resonance spectroscopy can be used to characterize glycosidic linkages, but this technique requires milligram amounts of material and cannot detect small amounts of coexisting isomers⁵. Mass spectrometry, on the other hand, can provide information on carbohydrate composition and connectivity for even small amounts of sample, but it cannot be used to distinguish between stereoisomers⁶. Here, we demonstrate that ion mobility–mass spectrometry—a method that separates molecules according to their mass, charge, size, and shape—can unambiguously identify carbohydrate linkage-isomers and stereoisomers. We analysed six synthetic carbohydrate isomers that differ in composition, connectivity, or configuration. Our data show that coexisting carbohydrate isomers can be identified, and relative concentrations of the minor isomer as low as 0.1 per cent can be detected. In addition, the analysis is rapid, and requires no derivatization and only small amounts of sample. These results indicate that ion mobility–mass spectrometry is an effective tool for the analysis of complex carbohydrates. This method could have an impact on the field of carbohydrate synthesis similar to that of the advent of high-performance liquid chromatography on the field of peptide assembly in the late 1970s.

The inherent structural diversity of glycans (carbohydrates that comprise a large number of monosaccharides, linked by glycosidic bonds) poses a major analytical challenge to all aspects of the glycosciences^{7,8}, and is one reason why glycomics lags behind the advances that have been made in genomics⁹ and proteomics¹⁰. The structure of a glycan is described by its composition, connectivity, and configuration (Fig. 1). The composition (Fig. 1a) is defined by its monosaccharides, the basic building blocks of oligosaccharides. These building blocks are often stereoisomers that differ only in their stereochemistry at one particular carbon atom, as in the case of glucose (Glc) and galactose (Gal). Each monosaccharide contains multiple hydroxyl groups that can be a point of attachment for a glycosidic bond with the next building block. Thus, unlike oligonucleotides and proteins, carbohydrates are not necessarily linear, but rather can be branched structures with diverse regiochemistry (that is, with linkages between different hydroxyl groups; Fig. 1b). In addition, a new stereocentre emerges when a glycosidic bond is formed, because two monosaccharides can be connected in two different configurations (Fig. 1c). These α - and β -anomers are stereoisomers, even though the connectivity is identical. Anomers are diastereomers, not enantiomers, meaning that they differ in at least one, but not all, of their stereocentres; consequently, anomers may differ in their size and properties.

When synthesizing oligosaccharides, managing different compositions is straightforward, because specific building blocks are added stepwise to generate the desired structure¹¹. Protective groups are used to define the connectivity by allowing the selective unveiling of specific hydroxyl groups¹², thus providing regiocontrol. In contrast, configurational control during glycosidic bond formation is the central challenge for chemical synthesis¹¹. *Trans*-glycoside formation is aided by the use of participating protecting groups. *Cis*-glycoside formation, however, cannot rely on participation, and anomeric mixtures are frequently obtained¹³.

The analysis of complex synthetic glycans is key to quality control but remains a major challenge. Glycan structure is typically ascertained by a combination of nuclear magnetic resonance spectroscopy (NMR) and mass spectrometry. Measuring a mass-to-charge ratio (m/z) with mass spectrometry is fast, requires very little sample and provides precise, high-resolution data about the sample composition. Detailed information regarding connectivity can be obtained following derivatization (chemical modification) and/or elaborate tandem mass spectrometry analysis^{6,14–16}. Nevertheless, with mass spectrometry it is not possible to analyse stereoisomers, since they generally cannot be distinguished from each other because of their identical atomic composition and mass. NMR experiments serve best to determine the configurational information of carbohydrates, but require large amounts of sample and are time-consuming; moreover, the resulting spectra are cumbersome to interpret when different stereoisomers need to be distinguished. In addition, the relative detection limit of

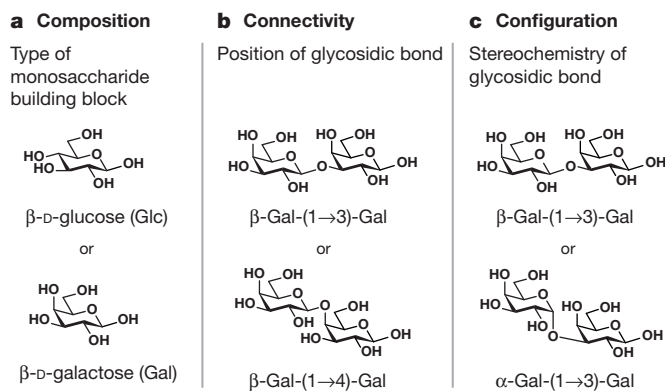


Figure 1 | Structural features of complex carbohydrates. **a**, The composition of a carbohydrate is defined by its monosaccharide content. Monosaccharide building blocks are often isomers, as shown for glucose (Glc) and galactose (Gal), which differ only in their C4 stereochemistry. **b**, Because of the many possible functional groups, the formation of a new glycosidic bond can occur at several positions, resulting in different connectivities, such as those shown here. **c**, Each glycosidic linkage is a new stereocentre that can have either α - or β -configuration.

¹Fritz Haber Institute of the Max Planck Society, Faradayweg 4–6, 14195 Berlin, Germany. ²Institute for Chemistry and Biochemistry, Free University Berlin, Takustraße 3, 14195 Berlin, Germany. ³Max Planck Institute of Colloids and Interfaces, Department of Biomolecular Systems, Am Mühlenberg 1, 14476 Potsdam, Germany. ⁴Institute for Chemistry and Biochemistry, Free University Berlin, Arnimallee 22, 14195 Berlin, Germany.

*These authors contributed equally to this work.

3% to 5% for larger oligosaccharides in NMR experiments is rather poor. Liquid chromatography can help to differentiate configurational isomers, but an unambiguous identification of one isomer in the presence of another is often not possible either⁸.

A promising approach to overcoming these limitations is the combination of ion mobility spectrometry and mass spectrometry (IM–MS)^{17,18}. Here, carbohydrate ions are separated not only according to their mass and charge, but also on the basis of their size and shape. IM–MS measures the time that ions require to drift through a cell that is filled with an inert neutral gas such as helium or nitrogen, under the influence of a weak electric field. While drifting, compact ions undergo fewer collisions with the gas than more extended ions, and therefore traverse the cell faster. This principle can allow for the separation of species with identical mass but different structure. The sample and time requirements of IM–MS are similar to those of a conventional mass spectrometry experiment, so the additional information is obtained at no extra cost. Moreover, the measured drift time can be converted into an instrument-independent, rotationally averaged collision cross-section (CCS). IM–MS has already been proven to be of value in the structural analysis of proteins and their assemblies^{19,20}, and also showed promise in the separation of carbohydrate and glycopeptide isomers^{21–26}. Previous IM–MS studies of regioisomers; data regarding the equally important compositional and configurational isomers is still lacking.

Here, we illustrate the utility of IM–MS for the in-depth structural analysis of carbohydrates by systematically investigating all types of isomerism simultaneously within one consistent and comparable set of

compounds. Six trisaccharide isomers (Fig. 2a) that, owing to their similarity in structure, are difficult to distinguish using established techniques were prepared using automated glycan assembly (Extended Data Fig. 1)^{11,27}. The six glycans share the reducing-end lactose motif Gal(β1→4)Glc, and an aminoalkyl linker placed during automated synthesis for conjugation to carrier proteins or array surfaces. The non-reducing-end moiety was varied, to generate isomers that differ in composition, connectivity, or configuration. Each of the glycan pairs **1** + **2** and **4** + **5** share the same regiochemistry and stereochemistry of their glycosidic linkages, but differ in their composition. On the other hand, the trisaccharide pairs **1** + **4**, **2** + **5**, and **3** + **6** are connectivity isomers, where the terminal building block is connected through either a 1→3 or a 1→4 glycosidic linkage. Finally, the glycan pairs **2** + **3** and **5** + **6** are configurational isomers that differ in the stereochemistry of the terminal glycosidic linkage. We analysed all six carbohydrates separately as both positively and negatively charged ions, using a commercially available hybrid IM–MS instrument (see Methods)²⁸. Although most previous studies focused on positive-ion adducts^{21–26}, we observed the most notable drift-time differences for deprotonated $[M-H]^-$ ions (Fig. 2b and Extended Data Fig. 3). As a result, we achieve a higher separation capability, and therefore these results will be used for the analysis below. The drift times of the individual sugars were further converted into CCSs (Extended Data Table 2) using a previously reported calibration protocol to enable comparison^{29,30}.

A comparison of the drift times and CCSs of the six trisaccharides revealed both similarities and differences (Fig. 2b). The compositional isomers **1** and **2** exhibit drift times and CCSs that are almost identical

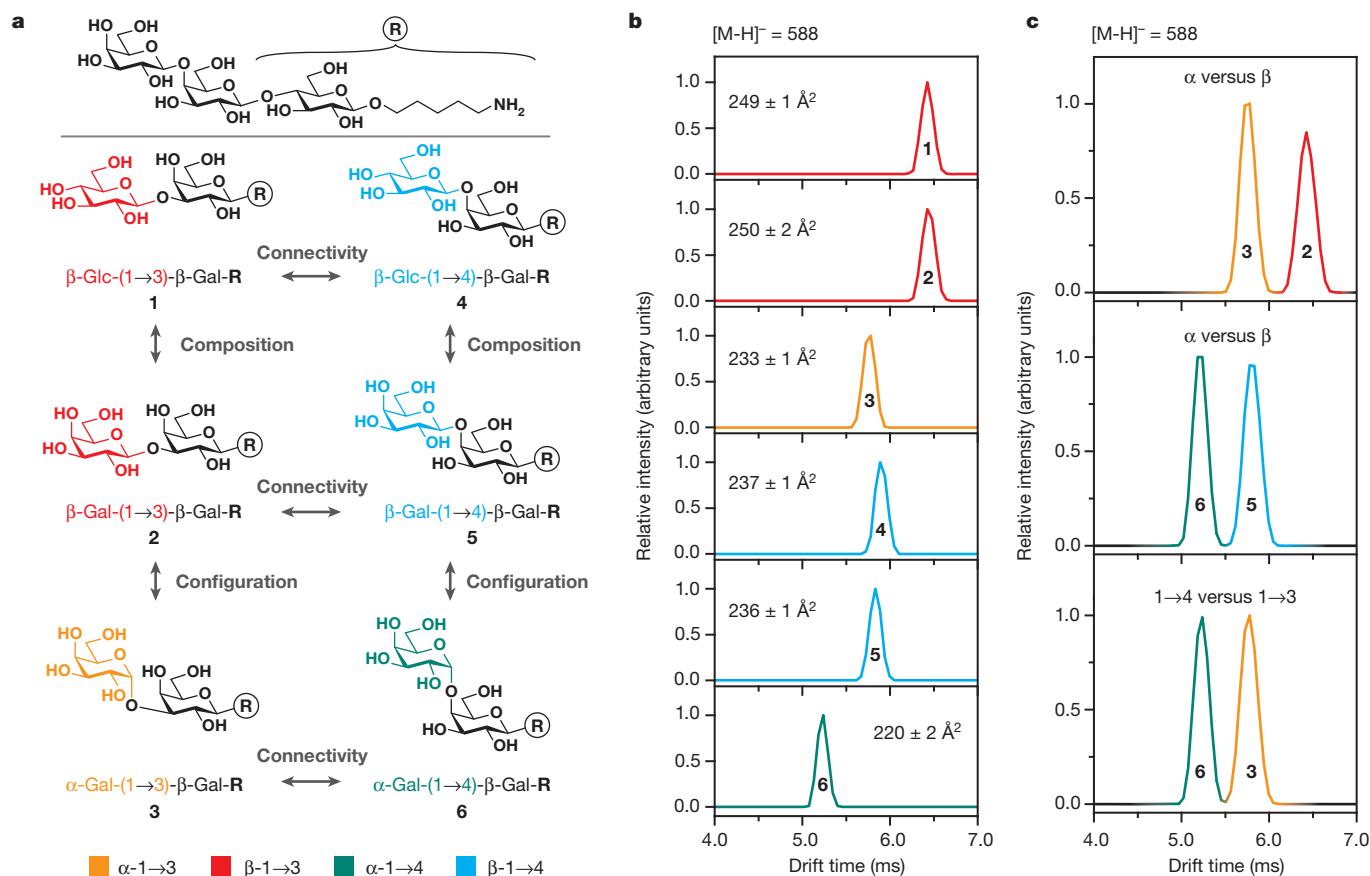


Figure 2 | Structure and IM–MS data of trisaccharides 1–6. **a**, The synthetic trisaccharides **1–6** share the same disaccharide core, and differ merely in the composition, connectivity, or configuration of the last monosaccharide building block. **b**, IM–MS drift-time distributions (also known as arrival-time distributions) for trisaccharides **1–6** as $[M-H]^-$ ions. The values in Å²

correspond to the estimated CCSs in the drift gas nitrogen and represent averages of three independent measurements. Although compositional isomers cannot be distinguished, connectivity and configurational isomers are clearly identified on basis of their CCSs. **c**, IM–MS drift-time distributions of isomeric mixtures show baseline separation between linkage- and stereoisomers.

to each other; the same is true of compositional isomers 4 and 5. This observation is not surprising, given that the respective trisaccharide pairs differ only in the orientation of one hydroxyl group, at the C4 carbon atom. Such a minimal structural difference is not expected to result in a notable difference in CCS. However, the composition of carbohydrates is easily controlled during automated synthesis, because either glucose or galactose building blocks are being used. Thus, compositional differentiation is not essential for the quality control of glycan synthesis.

In contrast to compositional isomers, regioisomers (1 + 4, 2 + 5, and 3 + 6) can be distinguished readily from each other on the basis of their drift times and CCSs. Here, the trisaccharides containing 1→3 glycosidic linkages exhibit larger CCSs than their 1→4-linked counterparts. Analytically, however, the most striking observations resulted from comparison of the 2 + 3 and 5 + 6 configurational isomers. α -Linked trisaccharides 3 and 6 adopt a more compact structure than do the corresponding β -linked molecules, and both sets of anomers can be clearly differentiated. While trisaccharides that differ in both regiochemistry and stereochemistry (for example, 3 and 5) exhibit similar CCSs in their respective deprotonated states, they can be distinguished in IM-MS as chloride adducts (Extended Data Fig. 3). In addition, deprotonated trisaccharide fragments generated from larger oligosaccharides exhibit highly diagnostic CCS values identical to those of their intact trisaccharide counterparts (Extended Data Fig. 4).

Although regiocontrol is well established during chemical oligosaccharide assembly, the formation of mixtures of stereoisomers is common when *cis*-glycosides are installed. The characterization and quality control of synthetic oligosaccharides would therefore benefit greatly from the ability to separate and identify different isoforms. We

systematically analysed mixtures of connectivity and configuration isomers by IM-MS (Fig. 2c). Strikingly, the linkage isomers (1→3 versus 1→4), as well as the α - and β -anomers, are fully baseline-separated (their peaks do not overlap). A similar quality of separation was also obtained in the quality control of a crude product mixture, where small amounts of an unintended by-product were clearly identifiable (Extended Data Fig. 6).

This encouraging result raises the question of whether an isomeric impurity can be not only identified qualitatively, but also determined quantitatively by IM-MS. We carried out experiments to address this question, where trisaccharide 2 was kept at a constant concentration, while the content of the corresponding anomer 3 was gradually reduced to mimic different percentages of a typical synthetic impurity (Fig. 3a). As expected, the intensity of the IM-MS peak of trisaccharide 3 gradually declines with decreasing concentration. An impurity with a relative concentration as little as 1% is still clearly visible and exhibits a well resolved and well shaped peak; however, although a relative content of 0.1% can still be identified qualitatively, it is close to the detection limit (see Methods and Extended Data Fig. 7b). To visualize better the large range of intensities, it is common to plot IM-MS data as a drift plot, with the drift time on the *x*-axis, *m/z* on the *y*-axis, and a logarithmic intensity scale (Fig. 3b). Here, a relative concentration of 0.1% is perfectly visible even without further magnification. We also tested the linearity of the IM-MS intensity for a broad range of concentrations with mixtures of anomers 2 and 3 (Fig. 3c). For that purpose, the relative peak area of the IM-MS signal of anomer 3 was plotted against the corresponding relative concentration. A value of 0.5 indicates an equal content of 2 and 3, while values of 0 and 1 are expected for the pure oligosaccharides, respectively. Remarkably, the

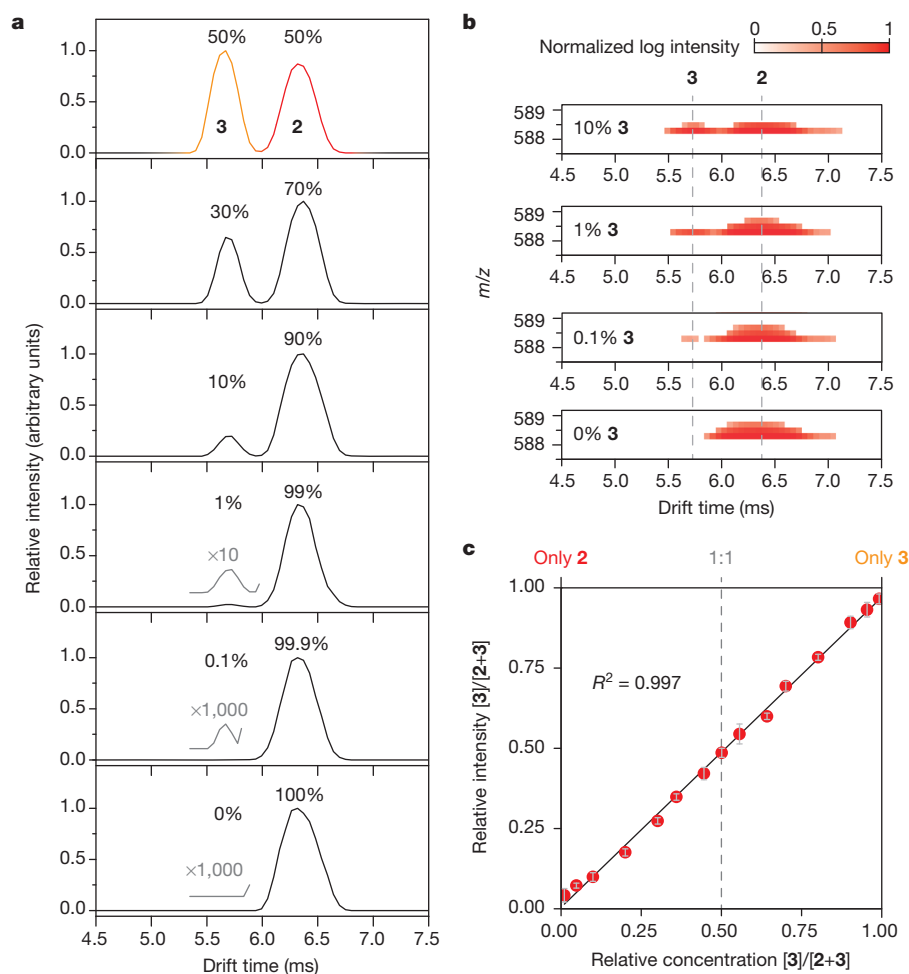


Figure 3 | Relative quantification of configurational trisaccharide isomers. Mixtures of the configurational isomers 2 and 3 were measured using IM-MS. **a**, The amount of isomer 2 was kept constant, while isomer 3 was diluted to yield relative concentrations of 50%, 30%, 10%, 1%, 0.1%, and 0%. Minor components with relative concentrations as low as 0.1% can still be qualitatively detected. The grey traces are magnified by the values shown. **b**, Three-dimensional plot showing the separation of anomers 2 and 3. The intensity is plotted using a logarithmic scale and impurities of 0.1% can be clearly identified without magnification. **c**, Plot of the relative IM-MS intensity of isomer 3 against the corresponding relative concentration, to illustrate the dynamic range of the method. A value of 0.50 represents equal amounts of isomer 2 and 3, while values of 0 and 1 indicate the presence of only isomer 2 or isomer 3, respectively. The grey error bars correspond to the double standard deviation observed for three independent replicates.

plot is strictly linear over the entire range of relative concentrations from 0.01 to 0.99, and very little deviation between different replicates is observed (for details see Methods and Extended Data Table 3). As a result, IM-MS can be used to estimate the relative content of a minor impurity when the investigated compounds are, like anomers, similar in structure and ionization efficiency.

In summary, we demonstrate that IM-MS is a powerful tool for the structural analysis and quality control of carbohydrates. Connectivity and configurational isomers can be separated efficiently with baseline resolution, especially when deprotonated ions are used, and the relative content of isomeric impurities can be determined quickly and easily (Extended Data Fig. 6). To our knowledge, no other experimental technique can provide the same structural information as quickly and with such minimal sample consumption. Larger glycans are known to separate less efficiently in IM-MS (Extended Data Fig. 5). However, their gas-phase fragments are similar to the oligosaccharides described here (Extended Data Fig. 4) and can therefore serve as diagnostic features for quality control and sequencing.

The full benefit of this method will become apparent once CCS data for carbohydrates and carbohydrate fragments, derived from synthetic and biological sources, are deposited in databases. These reference data will be essential for the quick and unambiguous identification of unknown compounds. The existence of commercially available mass spectrometers will enable IM-MS to become a routine technique for non-specialists or in automated analyses. IM-MS has the potential to fundamentally change quality control and analysis in carbohydrate chemistry.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 5 March; accepted 29 July 2015.

Published online 30 September 2015.

- Dwek, R. A. Glycobiology: toward understanding the function of sugars. *Chem. Rev.* **96**, 683–720 (1996).
- Molinari, M. *N*-glycan structure dictates extension of protein folding or onset of disposal. *Nature Chem. Biol.* **3**, 313–320 (2007).
- Varki, A. Sialic acids in human health and disease. *Trends Mol. Med.* **14**, 351–360 (2008).
- Bertozzi, C. R. & Rabuka, D. in *Essentials of Glycobiology* (eds Varki, A. et al.) Ch. 2 (Cold Spring Harbor Laboratory Press, 2009).
- Duus, J. Ø., Gottfredsen, C. H. & Bock, K. Carbohydrate structural determination by NMR spectroscopy: modern methods and limitations. *Chem. Rev.* **100**, 4589–4614 (2000).
- Dell, A. & Morris, H. R. Glycoprotein structure determination by mass spectrometry. *Science* **291**, 2351–2356 (2001).
- Service, R. F. Looking for a sugar rush. *Science* **338**, 321–323 (2012).
- Mariño, K., Bones, J., Kattla, J. J. & Rudd, P. M. A systematic approach to protein glycosylation analysis: a path through the maze. *Nature Chem. Biol.* **6**, 713–723 (2010).
- Venter, J. C. et al. The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
- Domon, B. & Aebersold, R. Mass spectrometry and protein analysis. *Science* **312**, 212–217 (2006).
- Plante, O. J., Palmacci, E. R. & Seeberger, P. H. Automated solid-phase synthesis of oligosaccharides. *Science* **291**, 1523–1527 (2001).
- Wang, Z. et al. A general strategy for the chemoenzymatic synthesis of asymmetrically branched *N*-glycans. *Science* **341**, 379–383 (2013).
- Boltje, T. J., Buskas, T. & Boons, G.-J. Opportunities and challenges in synthetic oligosaccharide and glycoconjugate research. *Nature Chem.* **1**, 611–622 (2009).
- Prien, J. M., Ashline, D. J., Lapadula, A. J., Zhang, H. & Reinhold, V. N. The high mannose glycans from bovine ribonuclease B isomer characterization by ion trap MS. *J. Am. Soc. Mass Spectrom.* **20**, 539–556 (2009).
- Daikoku, S., Widmalm, G. & Kanie, O. Analysis of a series of isomeric oligosaccharides by energy-resolved mass spectrometry: a challenge on homobranch trisaccharides. *Rapid Commun. Mass Spectrom.* **23**, 3713–3719 (2009).
- Harvey, D. J. Fragmentation of negative ions from carbohydrates: part 1. Use of nitrate and other anionic adducts for the production of negative ion electrospray spectra from *N*-linked carbohydrates. *J. Am. Soc. Mass Spectrom.* **16**, 622–630 (2005).
- Bohrer, B. C., Merenbloom, S. I., Koeniger, S. L., Hilderbrand, A. E. & Clemmer, D. E. Biomolecule analysis by ion mobility spectrometry. *Annu. Rev. Anal. Chem.* **1**, 293–327 (2008).
- Utrecht, C., Rose, R. J., van Duijn, E., Lorenzen, K. & Heck, A. J. R. Ion mobility mass spectrometry of proteins and protein assemblies. *Chem. Soc. Rev.* **39**, 1633–1655 (2010).
- Ruotolo, B. T. et al. Evidence for macromolecular protein rings in the absence of bulk water. *Science* **310**, 1658–1661 (2005).
- Bleiholder, C., Dupuis, N. F., Wyttenbach, T. & Bowers, M. T. Ion mobility-mass spectrometry reveals a conformational conversion from random assembly to β -sheet in amyloid fibril formation. *Nature Chem.* **3**, 172–177 (2011).
- Gabryelski, W. & Froese, K. L. Rapid and sensitive differentiation of anomers, linkage, and position isomers of disaccharides using high-field asymmetric waveform ion mobility spectrometry (FAIMS). *J. Am. Soc. Mass Spectrom.* **14**, 265–277 (2003).
- Plasencia, M. D., Isailovic, D., Merenbloom, S. I., Mechref, Y. & Clemmer, D. E. Resolving and assigning *N*-linked glycan structural isomers from ovalbumin by IMS-MS. *J. Am. Soc. Mass Spectrom.* **19**, 1706–1715 (2008).
- Zhu, M., Bendiak, B., Clowers, B. & Hill, H. H. Jr. Ion mobility-mass spectrometry analysis of isomeric carbohydrate precursor ions. *Anal. Bioanal. Chem.* **394**, 1853–1867 (2009).
- Williams, J. P. et al. Characterization of simple isomeric oligosaccharides and the rapid separation of glycan mixtures by ion mobility mass spectrometry. *Int. J. Mass Spectrom.* **298**, 119–127 (2010).
- Fenn, L. S. & McLean, J. A. Structural resolution of carbohydrate positional and structural isomers based on gas-phase ion mobility-mass spectrometry. *Phys. Chem. Chem. Phys.* **13**, 2196–2205 (2011).
- Both, P. et al. Discrimination of epimeric glycans and glycopeptides using IM-MS and its potential for carbohydrate sequencing. *Nature Chem.* **6**, 65–74 (2014).
- Kröck, L. et al. Streamlined access to conjugation-ready glycans by automated synthesis. *Chem. Sci.* **3**, 1617–1622 (2012).
- Pringle, S. D. et al. An investigation of the mobility separation of some peptide and protein ions using a new hybrid quadrupole/travelling wave IMS/oa-ToF instrument. *Int. J. Mass Spectrom.* **261**, 1–12 (2007).
- Pagel, K. & Harvey, D. J. Ion mobility-mass spectrometry of complex carbohydrates—collision cross sections of sodiated *N*-linked glycans. *Anal. Chem.* **85**, 5138–5145 (2013).
- Hofmann, J. et al. Estimating collision cross sections of negatively charged *N*-glycans using traveling wave ion mobility-mass spectrometry. *Anal. Chem.* **86**, 10789–10795 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank the Free University Berlin and the Max Planck Society for financial support. J.H. and K.P. thank G. von Helden, J.L.P. Benesch and W.B. Struwe for comments.

Author Contributions P.H.S. and K.P. designed the research; J.H. and H.S.H. performed the research. All authors analysed data and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to P.H.S. (peter.seeberger@mpikg.mpg.de) or K.P. (kevin.pagel@fu-berlin.de).

METHODS

Materials. All chemicals were reagent grade and used as supplied. Before use, molecular sieves were activated by heating under high vacuum. All reactions were performed in oven-dried glassware under an argon atmosphere, unless noted otherwise. *N,N*-dimethylformamide (DMF), dichloromethane (DCM), toluene and tetrahydrofuran (THF) were purified in a cycle-Tainer solvent delivery system.

Pre-automation steps. All elements such as the synthesizer, modules, linker-bound resin (**10**), and a set of building blocks (**11–19**) were prepared (see Supplementary Information for details).

Automated glycan assembly. All automated glycosylations were performed on an automated oligosaccharide synthesizer demonstrator unit using anhydrous solvents of the cycle-Tainer solvent delivery system. Oligosaccharides **1–9** were synthesized following the schemes in Extended Data Figs 1 and 2 using sequences I to V (Extended Data Table 1)^{11,27,31–33}. Detailed synthesis information and NMR data can be obtained from the Supplementary Information. To start the synthesis sequence, the resin was swollen in 2 ml DCM. The building blocks were coevaporated with toluene three times, dissolved in DCM under an argon atmosphere, and transferred into vials that were placed on the corresponding ports in the synthesizer. Reagents were dissolved in the corresponding solvents under an argon atmosphere in bottles that were placed on the corresponding ports in the synthesizer.

Module 1: acidic TMSOTf wash. The resin is washed three times with each of DMF, THF and DCM (each 2 ml for 25 s), and then once with 0.35 ml of a solution of trimethylsilyl trifluoromethanesulfonate (TMSOTf) in DCM at -20°C . The resin is swollen in 2 ml DCM and the temperature of the reaction vessel is adjusted to the set temperature, T_a .

Module 2: glycosylation using thioglycoside. For glycosylation, the DCM is drained and a solution of thioglycoside building block (5 equivalents in 1.0 ml DCM) is delivered to the reaction vessel. After the set temperature (T_a) is reached, the reaction starts with the addition of 1 ml *N*-iodosuccinimide (NIS, 5 equivalents in 1.0 ml DCM) and trifluoromethanesulfonic acid (TfOH, 0.1 equivalents in 1.0 ml DCM) solution. The glycosylation is performed for time t_1 at temperature T_a and for time t_2 at temperature T_i . After the reaction, the solution is drained and the resin is washed with DCM (six times with 2 ml for 15 s each). This procedure is repeated twice.

Module 3: Fmoc deprotection. The resin is washed with DMF (six times with 2 ml for 15 s) and swollen in 2 ml DMF; then the temperature of the reaction vessel is adjusted to 25°C . For fluorenylmethyloxycarbonyl (Fmoc) deprotection, the DMF is drained and 2 ml of a solution of 20% triethylamine in DMF is delivered to the reaction vessel. After 5 minutes, the reaction solution is collected in the fraction collector of the oligosaccharide synthesizer and 2 ml of a solution of 20% triethylamine in DMF is delivered to the resin. This procedure is repeated three times.

Module 4: glycosylation using phosphate. For glycosylation the DCM is drained and a solution of phosphate building block (5 equivalents in 1.0 ml DCM) is delivered to the reaction vessel. After the set temperature (T_a) is reached, the reaction starts with the addition of 1 ml TMSOTf solution. The glycosylation is performed for t_1 at T_a , and for t_2 at T_i . After the reaction the solution is drained and the resin is washed with DCM (six times with 2 ml for 15 s). This procedure is repeated twice.

Module 5: levulinoyl (lev) deprotection. The resin is washed with DCM (six times with 2 ml for 25 s) and swollen in 1.3 ml DCM; the temperature of the reaction vessel is adjusted to 25°C . For levulinoyl deprotection, 0.8 ml of the hydrazine hydrate solution is delivered into the reaction vessel. After 30 minutes the reaction solution is drained and the resin is washed with 0.2 M acetic acid in DCM and DCM (six times each with 2 ml for 25 s). The entire procedure is performed three times.

Purification of protected oligosaccharides. Following ultraviolet cleavage using the continuous-flow photoreactor³¹, the crude molecules were confirmed with matrix-assisted laser desorption/ionization (MALDI) and crude NMR (^1H , and heteronuclear single quantum coherence (HSQC) spectra, recorded on a Varian Mercury 400 (400 MHz) or 600 (600 MHz) spectrometer). In addition, the crude material was analysed by high-performance liquid chromatography (HPLC; Agilent 1100 series spectrometer; column: Luna 5 μm , silica 100 \AA (260×4.60 mm); flow rate 1 ml min⁻¹; eluents 5% DCM in hexane/5% DCM in ethyl acetate; gradient 20% (for 5 min), 60% (in 40 min), 100% (in 5 min); detection 280 nm, and evaporating light scattering detection (ELSD)). The samples were purified using preparative HPLC (Agilent 1200 series). The crude mixture was carefully dissolved in a minimum volume of DCM and 0.9 ml of 20% hexane in ethyl acetate, and injected for purification using preparative HPLC (column: Luna 5 μm , silica 260×10 mm; flow rate 5 ml min⁻¹; eluents 5% DCM in hexane/5% DCM in ethyl acetate; gradient 20% (for 5 min), 60% (in 40 min), 100% (in 5 min); detection 280 nm, and ELSD) to afford the fully protected target oligosaccharide.

Oligosaccharide deprotection and final purification. To the solution of the fully protected oligosaccharide in methanol (0.2 ml μmol^{-1} of oligosaccharide), 58 μl of 0.5 M sodium methoxide (NaOMe) solution (0.25 equivalents per acetyl of benzoyl group) in methanol was added at 40°C . The reaction was monitored by mass spectrometry until it was completed, then neutralized by 200 mg of Amberlite (400 mg per 100 μl NaOMe solution). After filtering off the suspension, the crude mixture was dissolved in methanol, ethyl acetate, and acetic acid (5:0.5:0.2, by volume), followed by the addition of 5% palladium on carbon (Pd/C) (50% by weight = Pd/oligosaccharide), and was then purged first with argon and then with hydrogen, and left to stir overnight at room temperature under balloon pressure. The reaction mixture was filtered through modified cellulose filter and washed with 20 ml water/methanol (9:1 in volume); the combined solution was evaporated under vacuum to provide the crude material. This material was analysed by reverse-phase HPLC (column Hypercarb (150×4.60 mm); flow rate 0.8 ml min⁻¹; eluents 0.1% formic acid in acetonitrile/0.1% formic acid in triple-distilled water; gradient 0% (for 10 min), 30% (in 30 min), 100% (in 5 min); detection ELSD). Subsequently the crude solution was purified by preparative reverse-phase HPLC (column Hypercarb, (150×10.00 mm); flow rate 3.6 ml min⁻¹; eluents 0.1% formic acid in acetonitrile/0.1% formic acid in triple-distilled water; gradient 0% (for 10 min), 30% (in 30 min), 100% (in 5 min); detection ELSD) to afford the unprotected oligosaccharide. All compounds were characterized by NMR, and high-resolution mass spectral analyses were performed using an Agilent 6210 electrospray ionization–time-of-flight spectrometer (Agilent Technologies). For details see Supplementary Information.

IM–MS. IM–MS experiments were performed on a travelling-wave quadrupole/ion mobility/orthogonal acceleration time-of-flight mass spectrometer, Synapt G2-S HDMS (Waters Corporation)²⁸, which was mass-calibrated before measurements using a solution of caesium iodide (100 mg ml⁻¹). IM–MS data analysis was performed using MassLynx 4.1, DriftScope 2.4 (Waters Corporation), and OriginPro 8.5 (OriginLab Corporation) software. For IM–MS analysis, compounds **1–9** and the crude mixture **5/30** were each dissolved in water/methanol (1:1 by volume) to a concentration of $1\text{--}10 \mu\text{mol l}^{-1}$. A nano-electrospray ionization source was used to ionize 3–5 μl of sample from platinum–palladium-coated borosilicate capillaries prepared in-house. Typical settings were: source temperature, 20°C ; needle voltage, 0.8 kV; sample cone voltage, 25 V; cone gas, off. The ion mobility parameters were optimized to achieve maximum resolution without excessive heating of the ions upon injection into the ion mobility cell. Values were: trap gas flow, 2 ml min⁻¹; helium cell gas flow, 180 ml min⁻¹; ion mobility gas flow, 90 ml min⁻¹; trap direct-current bias, 35 V; ion mobility wave velocity, 800 m s⁻¹; ion mobility wave height, 40 V. For MS/MS experiments the trap collision energy was increased to 30–60 V.

IM–MS spectra of each individual carbohydrate and three trisaccharide mixtures (**6 + 3**, **3 + 2** and **5 + 6**) were recorded separately in positive- and negative-ion mode. Drift-time distributions were extracted from raw data using MassLynx and drift times were determined manually via Gaussian fitting using Origin 8.5. For the measurement of the individual carbohydrates, the m/z signal intensity was kept at approximately 10^3 counts per second to avoid saturation and subsequent broadening of the corresponding drift peak (for an example see Extended Data Fig. 7). To avoid discrimination of a minor component, an average signal intensity of 10^4 counts per second was used for the semiquantitative assessment of mixtures (Extended Data Fig. 7b). Under these conditions, minor components with relative concentrations below 1% can still be detected qualitatively, but a semiquantitative assessment is no longer possible. For unknown mixtures, we therefore suggest acquiring data at both high- and low-intensity settings when possible. At high intensity, minor components with relative concentrations below 1% can be qualitatively detected, while the low-intensity case typically yields a better ion mobility resolution and enables a semiquantitative assessment (Extended Data Fig. 7). In addition, an acquisition at different intensity settings can help to evaluate mixtures in which the isomers cannot be fully resolved. For broad and inconclusive drift-time distributions, a comparison with neighbouring peaks of similar mass and charge can furthermore be used to distinguish between overlapping and saturated peaks²⁹.

CCS estimations were performed using an established protocol and dextran as calibrant (dextran1000, number average molecular weight 1,000; and dextran5000, number average molecular weight 5,000; Sigma Aldrich)^{29,30}. The calibration solution consisted of 0.1 mg ml⁻¹ dextran1000, 0.5 mg ml⁻¹ dextran5000, and 1 mM NaH_2PO_4 in water/methanol (1:1 by volume). The calibrant and each sample were measured on a travelling wave Synapt instrument at five wave velocities in both positive- and negative-ion mode. Drift times were extracted from raw data by fitting a Gaussian distribution to the drift-time distribution of each ion and corrected for their m/z -dependent flight time. CCS reference values³⁰ for dextran were corrected for charge and mass, and a logarithmic plot of corrected CCSs against corrected drift times was used as a calibration curve to estimate CCSs. One

calibration curve was generated for every wave velocity and each ion polarity. The resulting five estimated CCSs for each sample ion were averaged. These measurements were repeated three times and the averaged values for different ions are presented in Extended Data Table 2. The reported error corresponds to the standard deviation obtained for three independent replicates.

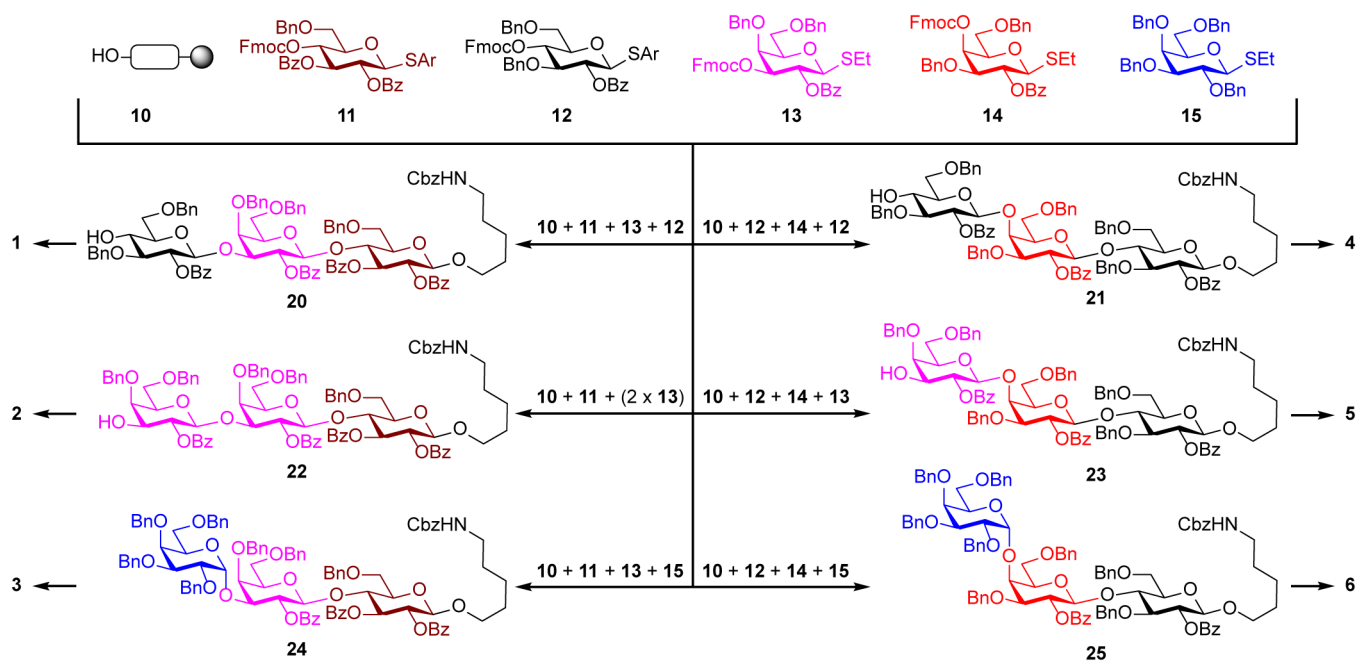
Semiquantitative analysis of trisaccharide mixtures. For the semiquantitative analysis of anomeric trisaccharide mixtures, a quantification experiment was performed using isomers **2** and **3**. Stock solutions of **2** and **3** with identical concentration were prepared in water/methanol (1:1 by volume). Each stock solution was diluted individually to yield relative concentrations of 80%, 56%, 43%, 25%, 11%, 5%, 1%, 0.1%, and 0.01%. The serial dilutions were used to obtain isomer mixtures with concentration ratios $[3]/[2] + [3]$ between 0 and 1 (see Extended Data Table 3). A value of 0.5 represents equal amounts of **2** and **3**, while 0 and 1 indicate the presence of only **2** or only **3**, respectively.

To achieve constant experimental conditions, we performed the semiquantitative analysis on a Synapt instrument equipped with an online nano-electrospray ionization source that was coupled to an ACQUITY ultraperformance liquid chromatography system (Waters). Settings were: eluents 0.1% formic acid in methanol/0.1% formic acid in water at a constant rate of 50%, flow rate $8 \mu\text{l min}^{-1}$, sample injection $10 \mu\text{l}$. Data were acquired in negative-ion mode with following settings: source temperature 80°C , needle voltage 2.7 kV; sample cone voltage 25 V, desolvation temperature 150°C , cone gas 0 l h^{-1} , nanoflow gas 1.3 bar, purge gas flow 500.0 ml h^{-1} . Ion mobility parameters were: trap gas flow, 0.4 ml min^{-1} ,

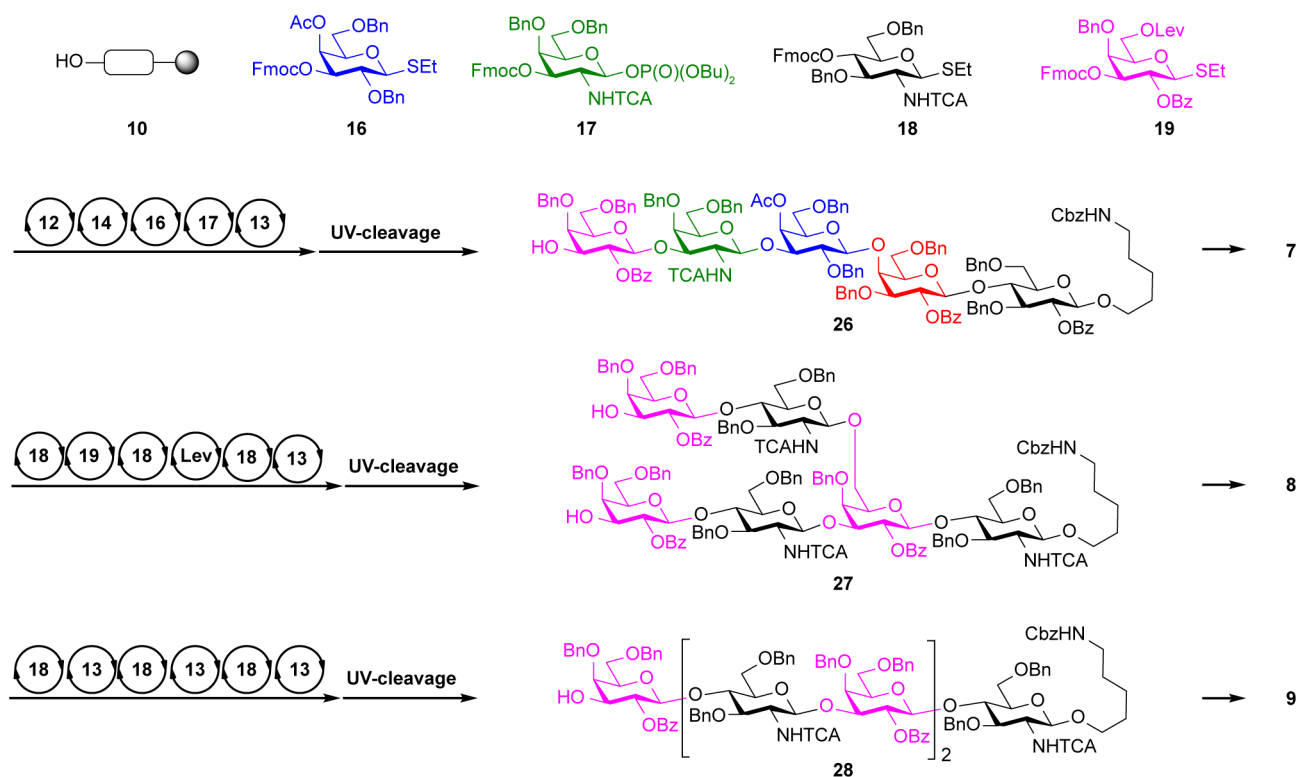
helium cell gas flow 180 ml min^{-1} , ion mobility gas flow 90 ml min^{-1} , trap direct-current bias 45 V, ion mobility wave velocity 800 m s^{-1} ; ion mobility wave height 40 V.

Extraction of the drift-time distribution of the 588.4 m/z ion showed two separate drift times, each of which corresponded to one of the two isomers. The area under the drift-time distribution is related to the concentration of the sample. Therefore, the theoretical concentration ratio was compared to the ratio of the drift-time peak areas (A) such that the relative intensity is $A(3)/A(2) + A(3)$ (Fig. 3c). A linear correlation was observed, demonstrating the semiquantification of one isomer in the presence of another, down to contents of 1% of the minor component. Relative concentrations between 1% and 0.1% were still qualitatively detectable, but a determination of the relative content was no longer possible owing to detector saturation caused by the major component.

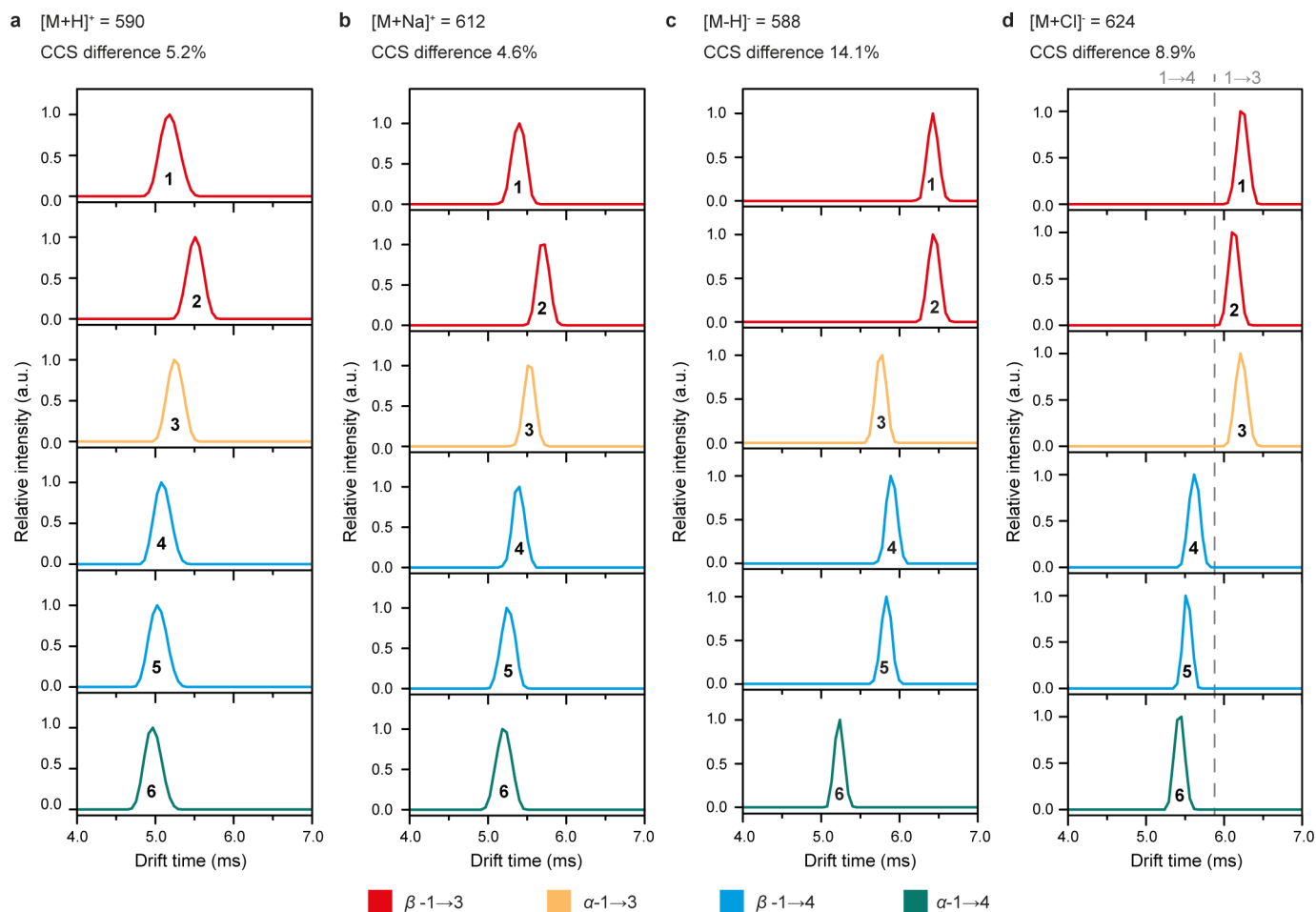
31. Eller, S., Collot, M., Yin, J., Hahm, H. S. & Seeberger, P. H. Automated solid-phase synthesis of chondroitin sulfate glycosaminoglycans. *Angew. Chem. Int. Ed.* **52**, 5858–5861 (2013).
32. Martin, C. E., Weishaupt, M. W. & Seeberger, P. H. Progress toward developing a carbohydrate-conjugate vaccine against *Clostridium difficile* ribotype 027: synthesis of the cell-surface polysaccharide PS-I repeating unit. *Chem. Commun.* **47**, 10260–10262 (2011).
33. Werz, D. B., Carstagner, B. & Seeberger, P. H. Automated synthesis of the tumor-associated carbohydrate antigens Gb-3 and Globo-H: incorporation of α -galactosidic linkages. *J. Am. Chem. Soc.* **129**, 2770–2771 (2007).



Extended Data Figure 1 | Automated synthesis of oligosaccharides 20–25. Ar, 2-methyl-5-tert-butylphenyl; Bn, benzyl; Bz, benzoyl; Cbz, carboxybenzyl; Et, ethyl; Fmoc, fluorenylmethyloxycarbonyl.

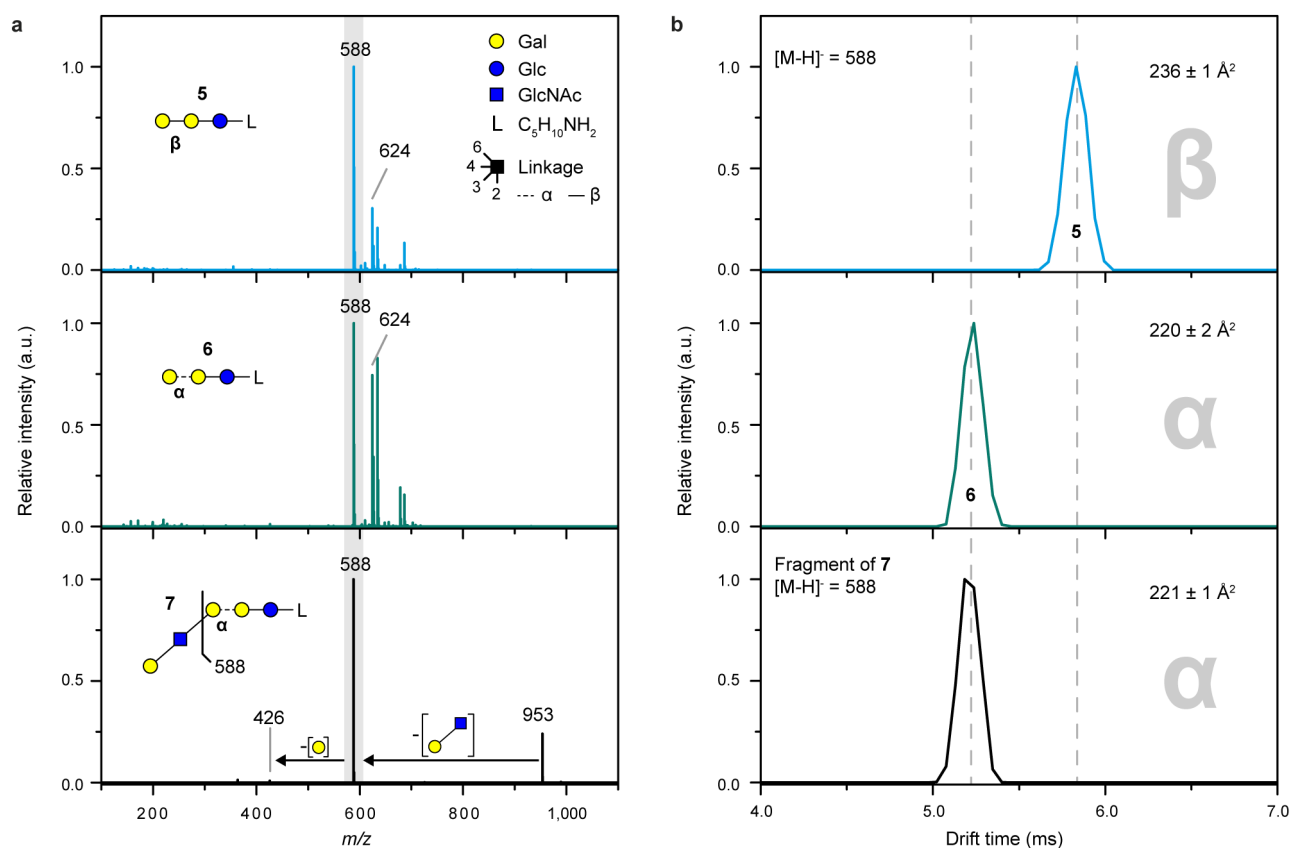


Extended Data Figure 2 | Automated synthesis of oligosaccharides 26–28. Ac, acetyl; Bn, benzyl; Bu, butyl; Bz, benzoyl; Cbz, carboxybenzyl; Et, ethyl; Fmoc, fluorenylmethyloxycarbonyl; lev, levulinoyl; TCA, trichloroacetimidate; UV, ultraviolet.



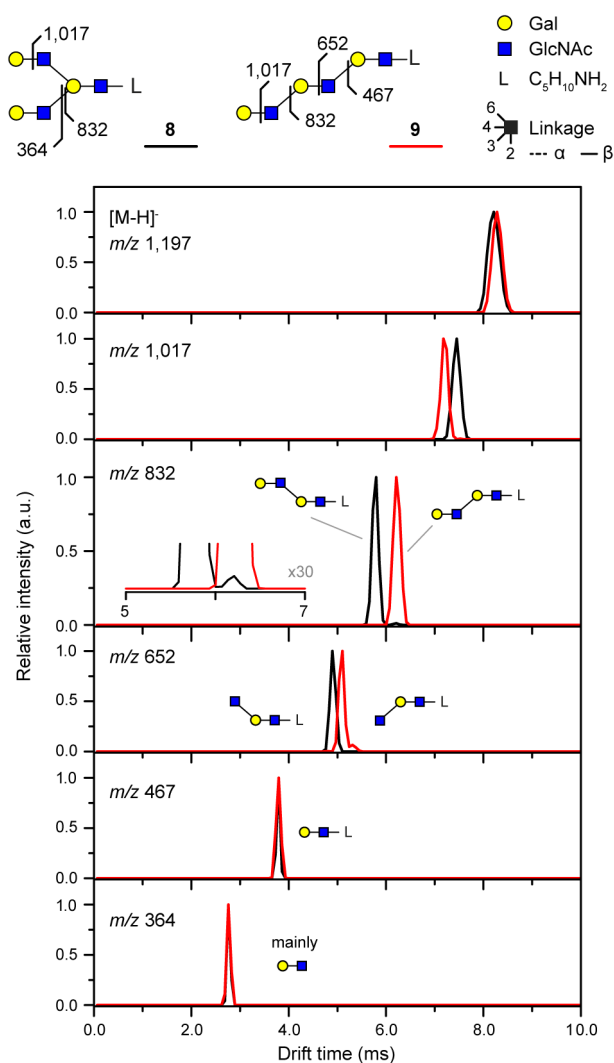
Extended Data Figure 3 | Drift-time distributions of trisaccharides 1–6 as different species in positive- and negative-ion mode. The CCS difference between the most compact and the most extended isomer of each set is given as a percentage. Small CCS differences are observed in positive-ion mode (a, b), which makes an unambiguous identification of the trisaccharides

difficult. The largest CCS differences are observed using deprotonated ions (c), allowing the identification of linkage isomers (for example, 3 + 6) and stereoisomers (for example, 2 + 3). A clear identification of regioisomers with a terminal 1→3 or 1→4 glycosidic bond can be obtained for chloride adducts (d).

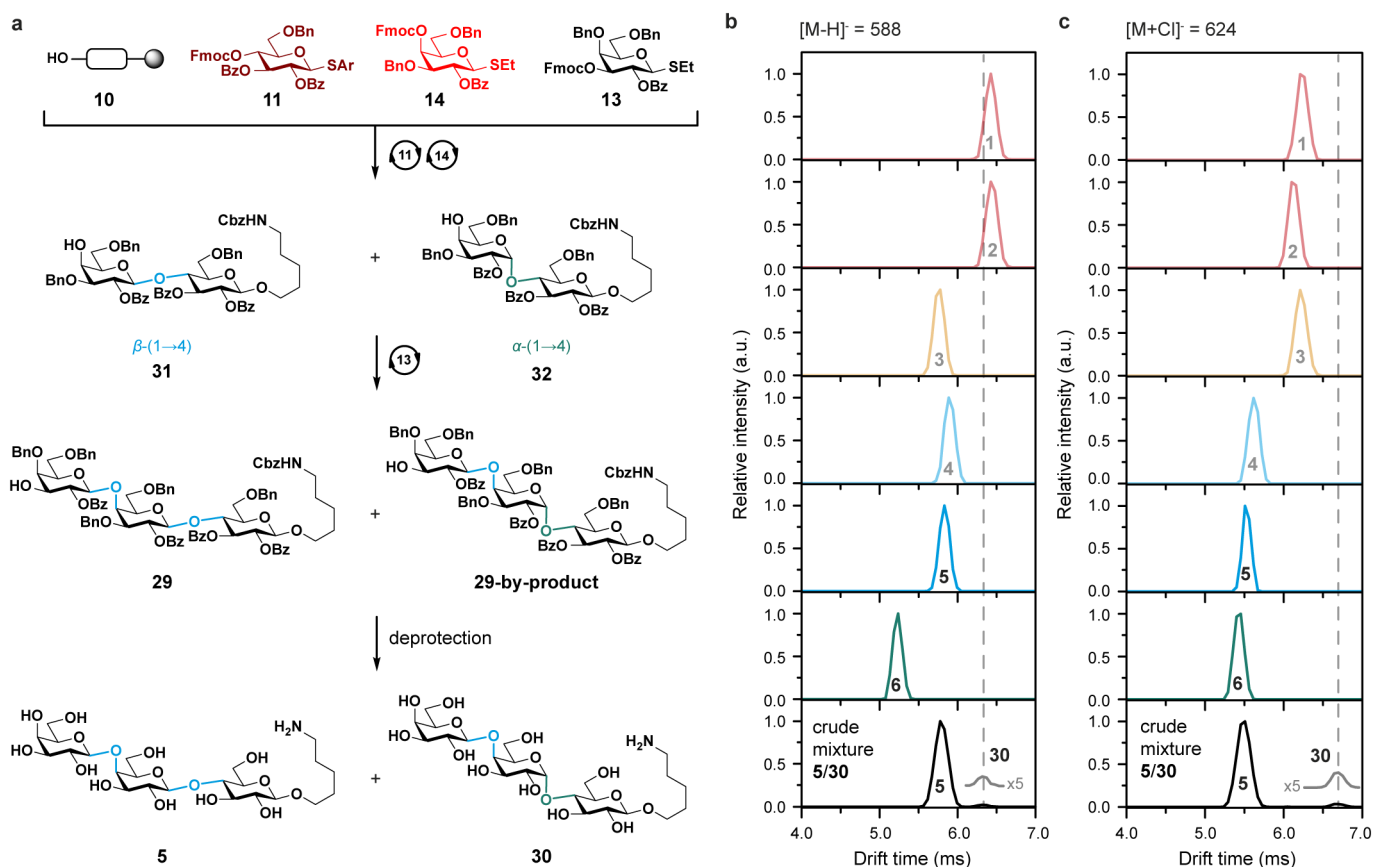


Extended Data Figure 4 | Comparison of drift times and CCSs of structurally similar precursor ions and fragments. **a**, Mass spectra of trisaccharides **5** and **6**, as well as a tandem MS spectrum of **7** (β -Gal-(1 \rightarrow 3)- β -GlcNAc-(1 \rightarrow 3)- α -Gal-(1 \rightarrow 4)- β -Gal-(1 \rightarrow 4)- β -Glc-L; L = $C_5H_{10}NH_2$) in negative-ion mode. The pentasaccharide **7** has the same core structure as the trisaccharide **6**. Collision-induced dissociation of deprotonated **7** consequently produces a fragment with the same mass as the deprotonated precursor

ion of **6**. **b**, Drift-time distributions of $[M-H]^- = 588$ ions. The collision-induced dissociation fragment arising from deprotonated **7** exhibits an drift time and CCS identical to those of the intact deprotonated trisaccharide **6**. This indicates that glycans and glycan fragments with identical structures also exhibit identical CCSs. Seen from a broader perspective, this highlights the exceptional potential of negative-ion CCSs to be used as a diagnostic parameter for glycan sequencing.

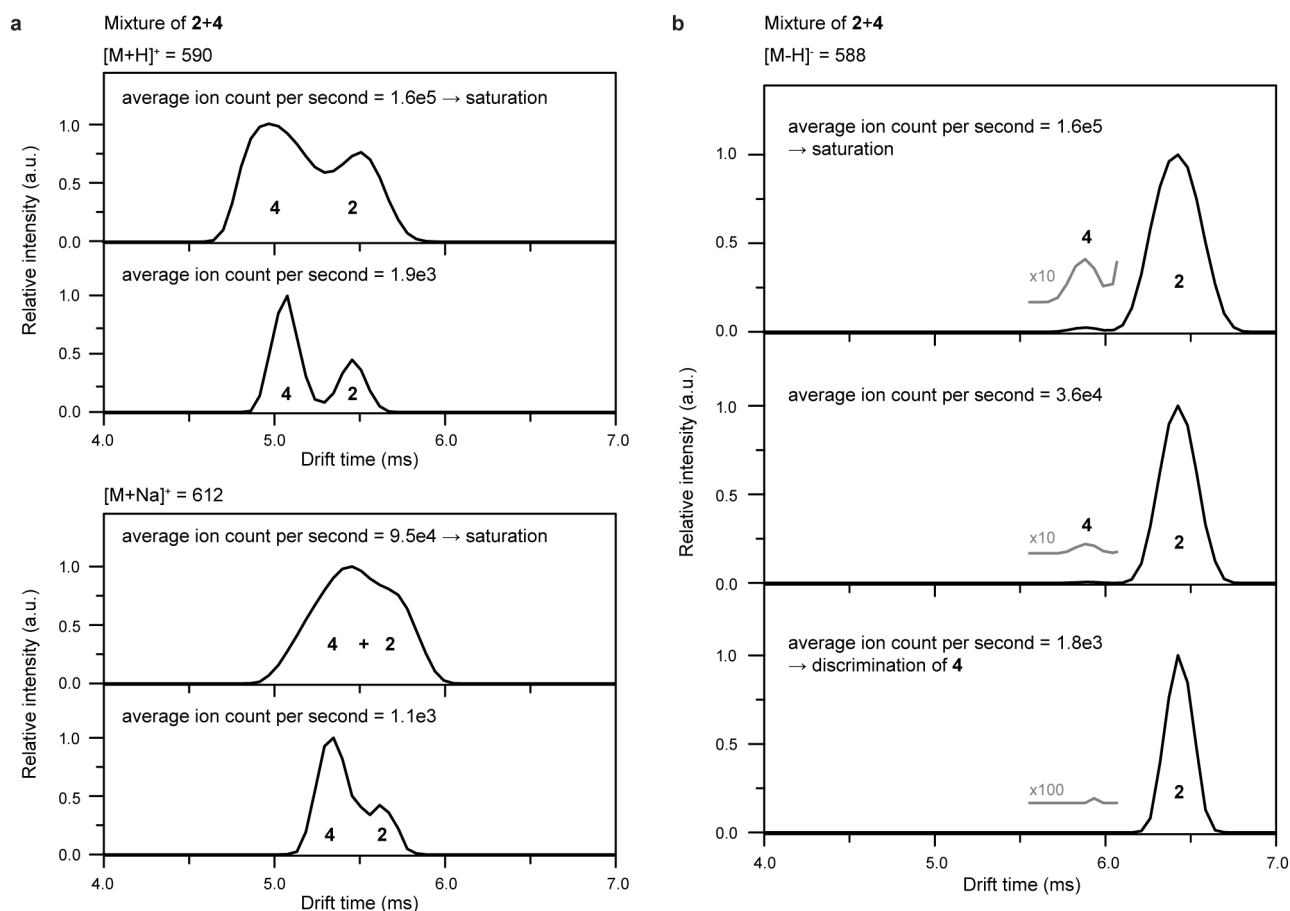


Extended Data Figure 5 | IM-MS differentiation and identification of the hexasaccharides **8 (black) and **9** (red).** As deprotonated ions, **8** and **9** show almost identical drift times and therefore cannot be distinguished. However, smaller collision-induced dissociation fragments containing five, four, or three monosaccharide building blocks (m/z 1,017, 832, and 652, respectively) exhibit highly diagnostic drift times. At m/z 832, a double peak is observed for the branched oligosaccharide **8** (inset, black trace), because two isomeric fragments are formed. Both fragments can be detected simultaneously using IM-MS, with cleavage at the 3-antenna being clearly preferred. The disaccharide fragments at m/z 467 and 364 are identical for **8** and **9** and consequently exhibit identical drift times.



Extended Data Figure 6 | Alternative synthesis of oligosaccharide 5 and corresponding IM-MS analysis. **a**, An alternative route for synthesizing 5 uses building block 11 instead of 12, which results in a mixture of the disaccharides 31 and 32 and subsequently in a mixture of trisaccharides 5 and 30. Neither the fully protected trisaccharides 29 and 29-by-product, nor the deprotected sugars 5 and 30, can be separated by HPLC. The formation of 29-by-product can be detected using NMR analysis, but a clear

structural assignment is not possible owing to the low relative concentration. **b**, $[M-H]^- = 588$ and **c**, $[M+Cl]^- = 624$ drift-time distributions of trisaccharides 1–6 compared to the drift time of the crude mixture consisting of 5 and 30 clearly reveal a content of about 5% by-product 30. In particular, the drift time of the chloride adduct of 30 is very diagnostic, because it differs considerably from the drift times of all other trisaccharides investigated here.



Extended Data Figure 7 | Correlation between signal intensity and ion mobility peak width in mixtures of 2 and 4. **a**, Drift-time distributions of [M + H]⁺ and [M + Na]⁺ ions at high (upper panels) and low (lower panels) signal intensity. The given average ion count per second corresponds to the signal detected for the major isotope peak. High signal intensities result in

peak broadening and reduced ion mobility resolution, whereas a considerably improved separation is achieved at lower intensity. **b**, Drift-time distributions of [M-H]⁻ ions from a mixture of <1% 4 and >99% 2. Measurements at high signal intensity can be used to qualitatively detect 4. At low intensity, however, 4 is discriminated and no signal can be detected.

Extended Data Table 1 | Sequences and conditions for automated oligosaccharide synthesis.

| sequence | module | details | condition |
|----------|--------|--|---|
| I | 1 | 2.5 eq. of TMSOTf solution | -20 °C, for 1 min |
| | 2 | 5 eq. building block for 11 , 12 and 18 , 5 eq. of NIS solution | $T_a = -30\text{ °C}$, $t_1 = 5\text{ min}$ $T_i = -30\text{ °C}$, $t_2 = 25\text{ min}$ |
| | 3 | Fmoc removal | r.t. for 5 min |
| II | 1 | 2.5 eq. of TMSOTf solution | -20 °C, for 1 min |
| | 2 | 5 eq. building block for 11 , 12 and 18 , 5 eq. of NIS solution | $T_a = -40\text{ °C}$, $t_1 = 5\text{ min}$ $T_i = -20\text{ °C}$, $t_2 = 25\text{ min}$ |
| | 3 | Fmoc removal | r.t. for 5 min |
| III | 1 | 2.5 eq. of TMSOTf solution | -20 °C, for 1 min |
| | 2 | 5 eq. of donor B, 5 eq. of NIS solution Fmoc removal | $T_a = -40\text{ °C}$, $t_1 = 5\text{ min}$ $T_i = -20\text{ °C}$, $t_2 = 25\text{ min}$ |
| IV | 1 | 2.5 eq. of TMSOTf solution | -20 °C, for 1 min |
| | 4 | 5 eq. building block 17 , 5 eq. of TMSOTf solution | $T_a = -40\text{ °C}$, $t_1 = 5\text{ min}$ $T_i = -20\text{ °C}$, $t_2 = 25\text{ min}$ |
| | 3 | Fmoc removal | r.t. for 5 min |
| V | 1 | 2.5 eq. of TMSOTf solution | -20 °C, for 1 min |
| | 5 | 5 eq. building block for 19 , 5 eq. of NIS solution | $T_a = -40\text{ °C}$, $t_1 = 5\text{ min}$ $T_i = -20\text{ °C}$, $t_2 = 25\text{ min}$ |
| | 5 | Lev removal | r.t. for 5 min |

Eq., equivalent; Fmoc, fluorenylmethyloxycarbonyl; Lev, levulinoyl; NIS, N-iodosuccinimide; r.t., room temperature; TMSOTf, trimethylsilyl trifluoromethanesulfonate.

Extended Data Table 2 | Estimated nitrogen CCSs ($^{TW}CCS_{N_2}$) for trisaccharides 1–6 and by-product 30.

| substance | ion | $^{TW}CCS_{N_2}$ in \AA^2 | STD in \AA^2 | ion | $^{TW}CCS_{N_2}$ in \AA^2 | STD in \AA^2 |
|-----------|--------------------|------------------------------------|-----------------------|---------------------|------------------------------------|-----------------------|
| 1 | [M+H] ⁺ | 231.9 | 0.4 | [M+Na] ⁺ | 236.2 | 0.6 |
| 2 | [M+H] ⁺ | 238.7 | 0.8 | [M+Na] ⁺ | 242.9 | 0.9 |
| 3 | [M+H] ⁺ | 233.6 | 0.7 | [M+Na] ⁺ | 239.6 | 0.6 |
| 4 | [M+H] ⁺ | 229.8 | 0.8 | [M+Na] ⁺ | 236.4 | 0.6 |
| 5 | [M+H] ⁺ | 228.9 | 0.6 | [M+Na] ⁺ | 233.6 | 0.7 |
| 6 | [M+H] ⁺ | 227.0 | 0.8 | [M+Na] ⁺ | 232.2 | 0.5 |
| 1 | [M-H] ⁻ | 249.4 | 1.1 | [M+Cl] ⁻ | 244.4 | 1.1 |
| 2 | [M-H] ⁻ | 249.8 | 1.5 | [M+Cl] ⁻ | 242.2 | 1.5 |
| 3 | [M-H] ⁻ | 233.2 | 1.3 | [M+Cl] ⁻ | 244.5 | 1.2 |
| 4 | [M-H] ⁻ | 237.4 | 0.9 | [M+Cl] ⁻ | 229.7 | 0.8 |
| 5 | [M-H] ⁻ | 235.6 | 1.0 | [M+Cl] ⁻ | 227.3 | 0.8 |
| 6 | [M-H] ⁻ | 219.9 | 1.6 | [M+Cl] ⁻ | 224.6 | 1.4 |
| 30 | [M-H] ⁻ | 248.4 | 0.3 | [M+Cl] ⁻ | 256.7 | 0.2 |

CCSs were estimated from travelling-wave (TW) measurements in nitrogen (N_2) using a previously described procedure^{29,30}. Each $^{TW}CCS_{N_2}$ is an average of three independent measurements with the corresponding standard deviation (STD).

Extended Data Table 3 | Relative concentrations of 2 and 3 in the investigated mixtures and their corresponding relative concentration ratio $x(3) = [3]/[3 + 2]$.

| | rel. conc. 3 | rel. conc. 2 | theoretical $x(3)$ | measured $\text{Int}_{\text{rel}}(3)$ | STD |
|-----|--------------|--------------|--------------------|---------------------------------------|-------|
| 1 | | 100 | 0.01 | 0.04 | 0.011 |
| 5 | | 100 | 0.05 | 0.07 | 0.004 |
| 11 | | 100 | 0.10 | 0.10 | 0.007 |
| 25 | | 100 | 0.20 | 0.18 | 0.005 |
| 43 | | 100 | 0.30 | 0.27 | 0.005 |
| 56 | | 100 | 0.36 | 0.35 | 0.005 |
| 80 | | 100 | 0.44 | 0.42 | 0.010 |
| 100 | | 100 | 0.50 | 0.49 | 0.007 |
| 100 | 80 | | 0.56 | 0.55 | 0.016 |
| 100 | 56 | | 0.64 | 0.60 | 0.005 |
| 100 | 43 | | 0.70 | 0.69 | 0.008 |
| 100 | 25 | | 0.80 | 0.78 | 0.005 |
| 100 | 11 | | 0.90 | 0.89 | 0.010 |
| 100 | 5 | | 0.95 | 0.93 | 0.012 |
| 100 | 1 | | 0.99 | 0.97 | 0.007 |

Measured relative intensities $\text{Int}_{\text{rel}}(3) = A(3)/A(2) + A(3)$ were calculated from the drift peak areas (A) of the deprotonated species $[M-H]^- = 588.4$. The standard deviation (STD) was obtained from three independent replicates.

Palaeomagnetic field intensity variations suggest Mesoproterozoic inner-core nucleation

A. J. Biggin¹, E. J. Piispa², L. J. Pesonen³, R. Holme¹, G. A. Paterson⁴, T. Veikkolainen³ & L. Tauxe⁵

The Earth's inner core grows by the freezing of liquid iron at its surface. The point in history at which this process initiated marks a step-change in the thermal evolution of the planet. Recent computational and experimental studies^{1–5} have presented radically differing estimates of the thermal conductivity of the Earth's core, resulting in estimates of the timing of inner-core nucleation ranging from less than half a billion to nearly two billion years ago. Recent inner-core nucleation (high thermal conductivity) requires high outer-core temperatures in the early Earth that complicate models of thermal evolution. The nucleation of the core leads to a different convective regime⁶ and potentially different magnetic field structures that produce an observable signal in the palaeomagnetic record and allow the date of inner-core nucleation to be estimated directly. Previous studies searching for this signature have been hampered by the paucity of palaeomagnetic intensity measurements, by the lack of an effective means of assessing their reliability, and by shorter-timescale geomagnetic variations. Here we examine results from an expanded Precambrian database of palaeomagnetic intensity measurements⁷ selected using a new set of reliability criteria⁸. Our analysis provides intensity-based support for the dominant dipolarity of the time-averaged Precambrian field, a crucial requirement for palaeomagnetic reconstructions of continents. We also present firm evidence for the existence of very long-term variations in geomagnetic strength. The most prominent and robust transition in the record is an increase in both average field strength and variability that is observed to occur between a billion and 1.5 billion years ago. This observation is most readily explained by the nucleation of the inner core occurring during this interval⁹; the timing would tend to favour a modest value of core thermal conductivity and supports a simple thermal evolution model for the Earth.

Palaeomagnetists have long sought to use data to constrain the thermal evolution of Earth through its influence on the geodynamo^{10–16}. In recent years, the quality and quantity of palaeomagnetic intensity (palaeointensity) measurements have increased substantially, allowing certain very-long-term variations in the Earth's dipole moment, and their possible causes, to be postulated. For example, a 'Proterozoic dipole low', extending from the earliest Proterozoic (about 2,450 million years (Myr) ago) to at least Cambrian (about 500 Myr ago) times was argued to reflect a weakened state of the geodynamo before the inner-core nucleation provided a substantial new power source⁷. More recently, the minimum of this weak-field interval was argued to be much earlier, at about 2,300–1,800 Myr ago¹⁷ and potentially linked to the existence of a dynamo generated in a basal magma ocean just above the outer core. Both of these studies suffered from limitations that we set out to address here: a shortage of measurement data in crucial time periods and an inability to demonstrate that claimed features were robust against sources of bias such as the intrinsic variability of the

magnetic field on timescales of tens of millions of years and less, and the varying reliability of the measurement data.

The present study uses a global compilation of 363 palaeointensity data (17% more than used by ref. 17 and with 41% more data in the interval 1,000–1,500 Myr ago) from the PINT database (<http://earth.liv.ac.uk/pint/>), all of which have been assigned palaeointensity quality (Q_{PI}) values⁸. These Q_{PI} values, applied at the palaeomagnetic site mean level, reflect the total number (with a maximum of nine, see Methods) of a set of individual criteria judged to have been met by a single palaeointensity estimate. For the purposes of this study, 43 estimates that had Q_{PI} values of 0 (or which were duplicates of other, higher-quality, data) were excluded, leaving 320 estimates from 36 studies (Supplementary Tables 1 and 2) for analysis.

Figure 1a and b shows the tendency of 118 of these palaeointensity results, selected because they were accompanied by suitable directional information (see Methods), to display a positive relationship between palaeointensity and palaeomagnetic inclination consistent with a dipole-dominated field. For $Q_{PI} \geq 1$ –5, all intensity data have significant positive Kendall rank correlations with inclination ($P \leq 0.0345$; see Methods). This result further supports the hypothesis that the geomagnetic field has been dipole-dominated for most of its history, which has previously only been investigated for the Precambrian era using directional data^{18,19}. The scatter about a dipole fit, as measured by the standard deviation about the expected intensity for a given inclination, decreases markedly as the minimum Q_{PI} value of the points is increased from 2 through to 4 (Fig. 1c), strengthening this observation and suggesting that Q_{PI} criteria are an effective means of assessing Precambrian-aged palaeointensity data.

The time evolution of the dipole moment was assessed using data sets with various minimum Q_{PI} values (Fig. 2, Extended Data Figs 1–4). A minimum Q_{PI} cut-off of 3 offers the optimal trade-off between misfit and quantity of data (Fig. 1c), but data sets produced using different cut-offs are also consistent with the findings detailed below (see Supplementary Table 3).

Dipole moment estimates are far from uniformly distributed through the assessed time period (500–3,500 Myr ago; Fig. 2). Within the more densely populated central time interval (1,000–2,800 Myr), virtual dipole moment and virtual axial dipole moment measurements (collectively referred to as VDMs henceforth) tend to be distributed into 'strips' of measurements made from units of individual igneous provinces with small differences in age. A large range of VDM measurements within a few million years or less is fully consistent with palaeomagnetic records from the past 2 Myr^{20,21}, supporting similar field behaviour throughout Earth history.

Similar to what is observed for the 0–200 Myr ago time period²², VDM measurements less than or equal to 50 ZAm² are ubiquitous in the Precambrian record (Fig. 2). In contrast, 'high' VDM measurements (greater than 50 ZAm²) are confined to time periods before

¹Department of Earth, Ocean and Ecological Sciences, University of Liverpool, Liverpool L69 7ZE, UK. ²Department of Geological and Mining Engineering and Sciences, Michigan Technological University, Houghton, 1400 Townsend Drive, Michigan 49931, USA. ³Department of Physics, Division of Materials Physics, PB 64, FI-00014 University of Helsinki, Helsinki, Finland. ⁴Key Laboratory of Earth and Planetary Physics, Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing 100029, China. ⁵Scripps Institution of Oceanography, University of California San Diego, La Jolla, California 92093-0220 USA.

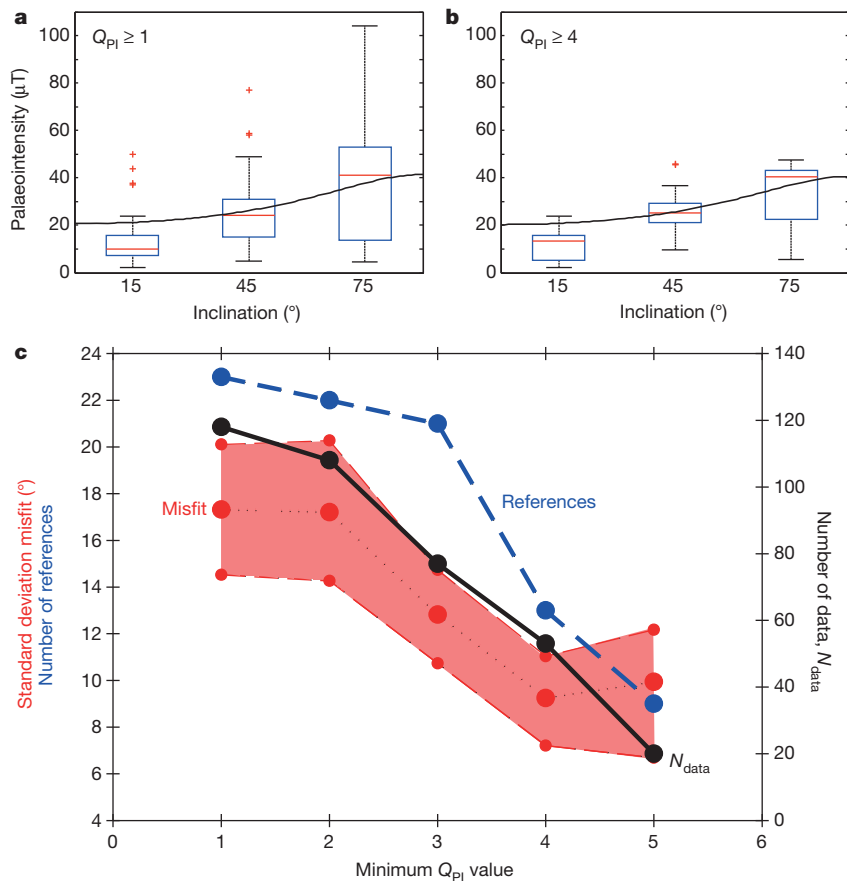


Figure 1 | Fits of palaeointensity data by minimum Q_{PI} value to palaeomagnetic inclination patterns predicted by a dipole field. **a, b,** Box-plots for all data in 30° inclination bins with minimum Q_{PI} values as shown. Horizontal lines are medians, boxes show the interquartile range (IQR), error bars show the full range excluding outliers (crosses) defined as being more than ± 1.5 IQR outside the box. **c,** Number of data N_{data} , number of references N_{ref} and model misfit (shading shows bootstrapped 95% uncertainties) versus minimum Q_{PI} value. Raw data are plotted in Extended Data Fig. 6a.

2,400 Myr ago (denoted 'Early') and after 1,300 Myr ago (denoted 'Late'). Some 48% of the estimates in these intervals (80 from a total of 166) are 'high' versus just 5% (2 from 41) in the intervening interval (denoted 'Mid'). Systematic bias from non-ideal rock magnetic behaviour or experimental procedures is very unlikely to be responsible for this disparity: 'Mid' interval measurements are sourced from 12 distinct studies (Supplementary Table 2) performed on a variety of lithologies (lavas, dykes, and plutons). Similarly, 'high' estimates from outside 'Mid' are sourced from 11 distinct studies (out of a total of 23) in the

two intervals, also from a variety of lithologies. Although the potential for biasing of palaeointensity estimates by poorly understood rock magnetic processes may remain even for results with high Q_{PI} values^{23–25}, this type of biasing is very unlikely to explain higher estimates being commonplace in certain parts of the Precambrian but nearly absent in other parts that are otherwise reasonably well represented.

Each data set ('Early', 'Mid' and 'Late') was analysed using non-parametric statistics (Fig. 3, Extended Data Fig. 5 and Supplemen-

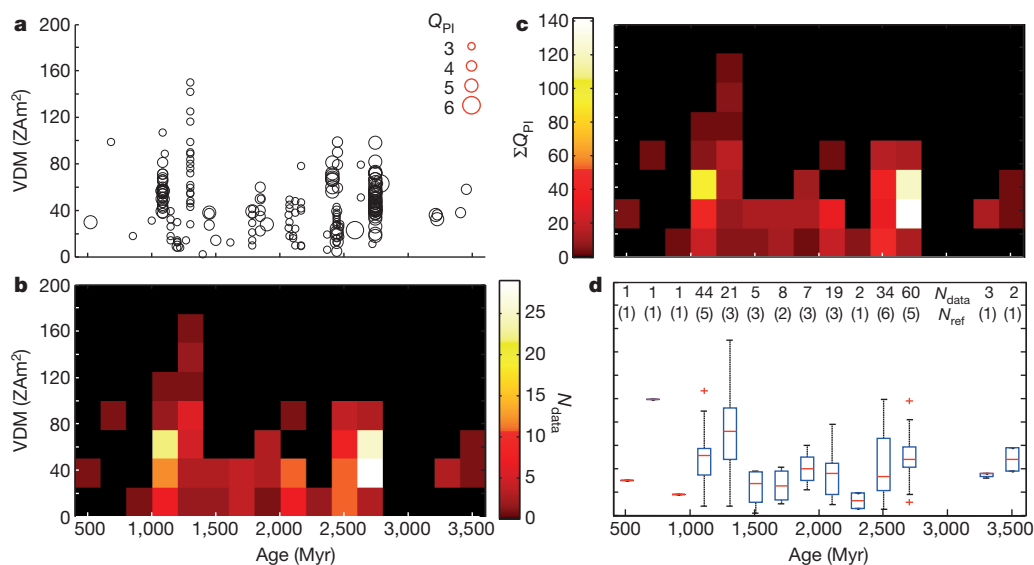


Figure 2 | Four different representations of VDM versus age for all data with $Q_{PI} \geq 3$. **a,** Bubble plot where size indicates Q_{PI} value. **b,** Density plot of number of measurements. **c,** Density plot of sum of Q_{PI} values. **d,** Box plot after

binning with an interval length of 200 Myr (number of data in each are given with the number of published studies in parentheses). See Fig. 1 caption for an explanation of the box plot.

tary Table 3). The distributions of VDMs with $Q_{PI} \geq 3$ in the ‘Late’ (median 54^{+3}_{-7} ZAm²) and ‘Early’ (median 44^{+6}_{-3} ZAm²) intervals are distinct beyond the 90% confidence limit ($P = 0.083$) according to the Kolmogorov–Smirnov test (Supplementary Table 3). ‘Mid’ has a median (30 ± 8 ZAm²) that is 32% lower than ‘Early’ and 44% lower than ‘Late’ and is distinct from both at a confidence limit of greater than 99.9%. The significance level of these disparities remains $>99\%$ using a minimum Q_{PI} cut-off of 4 and far exceeds this using cut-offs of 1 and 2 (Supplementary Table 3). A further resampling test (see Methods), incorporating quoted uncertainties in both the VDMs and their associated ages also produces significant results for a Q_{PI} cut-off of 3 or below (Extended Data Table 1).

To investigate whether the differences observed between our intervals could be explained by oversampling of geomagnetic variations occurring on timescales shorter than those which we are interested in here, we devised a tailored likelihood test (see Methods). This incorporates the effects of bias arising from large (factor of 3) and long-lasting (50 Myr, chosen to be longer than any known superchron) shifts in the time-averaged dipole that is probably due to variable mantle forcing²⁶. It also incorporates the effects of bias potentially caused by ‘normal’ secular variation on the clustering of measurements, derived from the same suite of igneous rocks, within periods of 200 thousand years (kyr).

Analysing the data with $Q_{PI} \geq 3$ (Extended Data Table 2) indicates that, as would be expected, substantial differences between the distributions of VDM data are much more likely to arise by chance when such sources of bias are considered. In particular, they could explain the differences observed between the VDM distributions produced from the ‘Early’ and ‘Mid’ intervals ($P = 0.187$). Nevertheless, the simultaneous observation of differences of the same magnitude as observed between time periods ‘Mid’ and ‘Late’ remains highly unlikely ($P = 0.015$) without appealing to either systematic measurement bias or some very-long-term evolution of the time-averaged dipole moment. Two studies^{27,28}, which were not fully represented in previous versions of the database, contribute 72% of the data within the ‘Late’ interval. Nevertheless, arbitrarily excluding all data from either one of these studies still yields a low likelihood ($P \leq 0.068$) that the differences observed exist by chance alone (Extended Data Table 2).

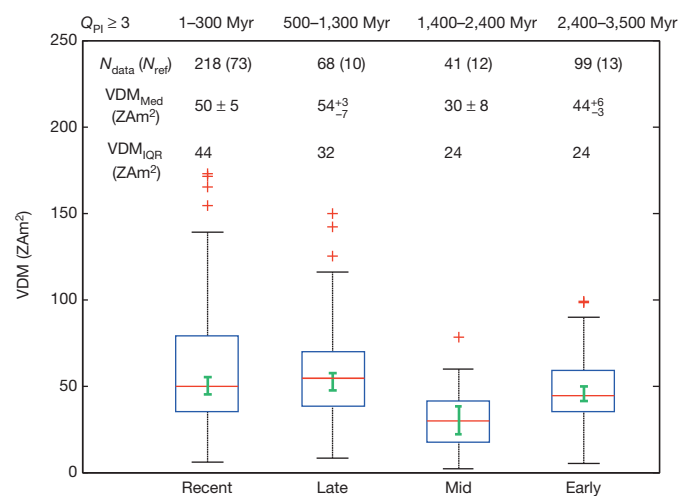


Figure 3 | Box-plot and summary statistics for different time intervals comprising VDM estimates with $Q_{PI} \geq 3$. N_{data} refers to the number of VDM estimates in each interval and N_{ref} refers to the number of published studies that these are drawn from. VDM_{med} and VDM_{IQR} refer, respectively, to the median and interquartile range of the VDM estimates within each interval. See Extended Data Table 1 for further information including the effect of varying the minimum Q_{PI} value. See Fig. 1 caption for an explanation of the box plot. Thick error bars indicate 95% confidence limits (from 10,000 bootstraps) on the median values.

Given that there is no good reason to suspect that both of them are biased high, we infer that the observed differences are robust.

Similarly, robust results (though with reduced levels of significance) are produced if we allow the long-term changes to increase to extreme factors of 6 and 12 (Extended Data Table 2). We conclude that our updated palaeointensity data set presents the first compelling evidence of geomagnetic intensity variations occurring on timescales longer than those that have previously been ascribed to mantle convection. Furthermore, a long-term increase in the time-averaged dipole moment very likely did occur at some time close to the end of our ‘Mid’ interval or near the beginning of our ‘Late’ time interval. Interestingly, the timing of this transition fits well with a recent finding²⁹ that the pattern of palaeomagnetic secular variation (based on purely directional data) shifted to a less stable state around 1,500 Myr ago.

Taken at face value, the record summarized in Fig. 2d (and Extended Data Figs 1d, 2d, 3d and 4d) indicates that there was a gradual decrease in dipole moment and its variability beginning in the late Archaean (about 2,500 Myr ago), which terminated with an abrupt increase in the Mesoproterozoic (about 1,300 Myr ago). A qualitatively similar pattern of dipole moment evolution through the Precambrian was predicted by a study⁹ employing a thermal evolution model coupled to the results of scaling analyses of numerical geodynamo models. Within the framework of this ‘low power’ end-member prediction (figure 11b in ref. 9), the gradual decrease in dipole moment would reflect the diminishing vigour of thermal convection caused by the secular cooling of the core; the subsequent sharp recovery at about 1,300 Myr ago would mark inner-core nucleation and the sudden commencement of much more efficient compositional convection. The thermal model in question⁹ predicted a somewhat earlier low-to-high transition (that is, age of inner-core nucleation) of about 1,800 Myr ago but we speculate that a slightly less extreme ‘low-mid power’ model could show good agreement with the record presented here.

A corollary of a Mesoproterozoic-age inner core and a conventional thermal history of the Earth is that the long-term dipole moment would probably have undergone only a small decrease since the onset of compositional convection⁹. Although intervening data are currently rare, a relative wealth of palaeointensity data are available in the interval 0–300 Myr ago, thus enabling a limited test of this hypothesis. A high-quality subset of these data (the ‘Recent’ interval, see Methods) yields a median VDM (50 ± 5 ZAm²) which is $\leq 10\%$ lower than that of the ‘Late’ interval (Fig. 3, Extended Data Fig. 5 and Supplementary Table 3). Similarly, a recent analysis²² of the last 200 Myr yielded a long-term median dipole moment of 42 ZAm² that is a maximum of 24% lower than the median values calculated for our ‘Late’ interval. Thus, the high values of dipole moment in our ‘Late’ interval are nearly matched by those within the ‘Recent’ interval, consistent with the very-long-term strength of the field decaying only marginally since inner-core nucleation. Our prediction is therefore supported by existing data and a more complete test will be possible in the future once the time period 300–1,000 Myr ago has been populated with reliable new palaeointensity measurements.

Our interpretation of the dipole moment record is not unequivocal because the implications of inner-core nucleation for the observable field at the Earth’s surface are not fully understood. Furthermore, some mantle-forced shift in core–mantle heat flow, lasting in excess of 50 Myr and perhaps related to the supercontinent cycle or secular mantle evolution, cannot be ruled out as causing a significant shift in geomagnetic behaviour during the Mesoproterozoic. Nevertheless, in the absence of rival thermal models making predictions similar to that which we have based our interpretation on, we argue that nucleation of the inner core in the Mesoproterozoic is at present the most likely explanation for the increase we have reported. Alternative candidates, potentially worth testing with models in the future, include increases in core–mantle heat flow resulting from the onset of whole-mantle convection (or even plate tectonics) and the first appearance of

post-perovskite (with associated elevated thermal diffusivity³⁰) at the base of the mantle.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 7 May; accepted 20 August 2015.

- Pozzo, M., Davies, C., Gubbins, D. & Alfè, D. Thermal and electrical conductivity of iron at Earth's core conditions. *Nature* **485**, 355–358 (2012).
- Seagle, C. T., Cottrell, E., Fei, Y. W., Hummer, D. R. & Prakapenka, V. B. Electrical and thermal transport properties of iron and iron-silicon alloy at high pressure. *Geophys. Res. Lett.* **40**, 5377–5381 (2013).
- Gomi, H. *et al.* The high conductivity of iron and thermal evolution of the Earth's core. *Phys. Earth Planet. Inter.* **224**, 88–103 (2013).
- de Koker, N., Steinle-Neumann, G. & Vlcek, V. Electrical resistivity and thermal conductivity of liquid Fe alloys at high P and T, and heat flux in Earth's core. *Proc. Natl Acad. Sci. USA* **109**, 4070–4073 (2012).
- Zhang, P., Cohen, R. E. & Haule, K. Effects of electron correlations on transport properties of iron at Earth's core conditions. *Nature* **517**, 605–607 (2015).
- Aubert, J., Tarduno, J. A. & Johnson, C. L. Observations and models of the long-term evolution of Earth's magnetic field. *Space Sci. Rev.* **155**, 337–370 (2010).
- Biggin, A. J., Strik, G. & Langereis, C. G. The intensity of the geomagnetic field in the late-Archaeon: new measurements and an analysis of the updated IAGA palaeointensity database. *Earth Planets Space* **61**, 9–22 (2009).
- Biggin, A. J. & Paterson, G. A. A new set of qualitative reliability criteria to aid inferences on palaeomagnetic dipole moment variations through geological time. *Frontiers Earth Sci.* **2**, 21–29 (2014).
- Aubert, J., Labrosse, S. & Poitou, C. Modelling the palaeo-evolution of the geodynamo. *Geophys. J. Int.* **179**, 1414–1428 (2009).
- Dunlop, D. J. & Yu, Y. in *Timescales of the Internal Geomagnetic Field* Vol. 145 of *Geophysical Monograph Series* (ed. Channell, J. E. T.) 85–100 (AGU, 2004).
- Prévot, M. & Perrin, M. Intensity of the Earth's magnetic-field since precambrian from Thellier-type paleointensity data and inferences on the thermal history of the core. *Geophys. J. Int.* **108**, 613–620 (1992).
- Biggin, A. J. & Thomas, D. N. Analysis of long-term variations in the geomagnetic poloidal field intensity and evaluation of their relationship with global geodynamics. *Geophys. J. Int.* **152**, 392–415 (2003).
- Valet, J. P. Time variations in geomagnetic intensity. *Rev. Geophys.* **41**, 1004, <http://dx.doi.org/10.1029/2001RG000104> (2003).
- Biggin, A. J., Strik, G. H. M. A. & Langereis, C. G. Evidence for a very-long-term trend in geomagnetic secular variation. *Nature Geosci.* **1**, 395–398 (2008).
- Macouin, M., Valet, J. P. & Besse, J. Long-term evolution of the geomagnetic dipole moment. *Phys. Earth Planet. Inter.* **147**, 239–246 (2004).
- Tauxe, L. & Yamazaki, T. in *Geomagnetism* Vol. 5 of *Treatise on Geophysics* (ed. Kono, M.) Ch. 13, 510–563 (Elsevier, 2007).
- Valet, J. P., Besse, J., Kumar, A., Vadakke-Chanat, S. & Philippe, E. The intensity of the geomagnetic field from 2.4 Ga old Indian dykes. *Geochem. Geophys. Geosyst.* **15**, 2426–2437 (2014).
- Evans, D. A. D. Proterozoic low orbital obliquity and axial-dipolar geomagnetic field from evaporite palaeolatitudes. *Nature* **444**, 51–55 (2006).
- Veikkolainen, T., Evans, D. A. D., Korhonen, K. & Pesonen, L. J. On the low-inclination bias of the Precambrian geomagnetic field. *Precamb. Res.* **244**, 23–32 (2014).
- Ziegler, L. B., Constable, C. G., Johnson, C. L. & Tauxe, L. PADM2M: a penalized maximum likelihood model of the 0–2 Ma palaeomagnetic axial dipole moment. *Geophys. J. Int.* **184**, 1069–1089 (2011).
- Valet, J. P., Meynadier, L. & Guyodo, Y. Geomagnetic dipole strength and reversal rate over the past two million years. *Nature* **435**, 802–805 (2005).
- Tauxe, L., Gee, J. S., Steiner, M. B. & Staudigel, H. Paleointensity results from the Jurassic: new constraints from submarine basaltic glasses of ODP Site 801C. *Geochem. Geophys. Geosyst.* **14**, 4718–4733 (2013).
- de Groot, L. V., Fabian, K., Bakelaar, I. A. & Dekkers, M. J. Magnetic force microscopy reveals meta-stable magnetic domain states that prevent reliable absolute palaeointensity experiments. *Nature Commun.* **5**, 4548, <http://dx.doi.org/10.1038/Ncomms5548> (2014).
- Smirnov, A. V. & Tarduno, J. A. Thermochemical remanent magnetization in Precambrian rocks: are we sure the geomagnetic field was weak? *J. Geophys. Res.* **110**, B06103 (2005).
- Tarduno, J. A. & Smirnov, A. V. in *Timescales of the Paleomagnetic Field* Vol. 145 *Geophysical Monograph Series* (eds Channell, J. E. T., Kent, D. V., Lowrie, W. & Meert, J. G.) 328 (AGU, 2004).
- Biggin, A. J. *et al.* Possible links between long-term geomagnetic variations and whole-mantle convection processes. *Nature Geosci.* **5**, 526–533 (2012).
- Kulakov, E. V., Smirnov, A. V. & Diehl, J. F. Absolute geomagnetic paleointensity as recorded by similar to 1.09 Ga Lake Shore Traps (Keweenaw Peninsula, Michigan). *Stud. Geophys. Geodaet.* **57**, 565–584 (2013).
- Thomas, N. An integrated rock magnetic approach to the selection or rejection of ancient basalt samples for paleointensity experiments. *Phys. Earth Planet. Inter.* **75**, 329–342 (1993).
- Veikkolainen, T. & Pesonen, L. J. Palaeosecular variation, field reversals and the stability of the geodynamo in the Precambrian. *Geophys. J. Int.* **199**, 1515–1526 (2014).
- Hunt, S. A. *et al.* On the increase in thermal diffusivity caused by the perovskite to post-perovskite phase transition and its implications for mantle dynamics. *Earth Planet. Sci. Lett.* **319–320**, 96–103 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank T. Torsvik for organising the 7th Nordic Supercontinents Meeting and acknowledge financial support for this from the European Research Council (ERC Advanced Grant 267631) and the Research Council of Norway through its Centres of Excellence funding scheme (CEED 223272). We also thank J. Rees and L. Waszek for discussions. A.J.B. acknowledges funding from a NERC standard grant (NE/H021043/1). G.A.P. acknowledges funding from an NSFC grant (41374072). L.T. acknowledges funding from an NSF grant (EAR 1345003).

Author Contributions A.J.B. designed the study. A.J.B., E.J.P., L.J.P. and T.V. assigned the Q_{PI} values. A.J.B., R.H., G.A.P., L.J.P., T.V. and L.T. wrote the paper. A.J.B., G.A.P. and T.V. analysed the data.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.J.B. (biggin@liv.ac.uk).

METHODS

Data selection. During the Nordic Supercontinents Workshop in Haraldvangen, Norway, in October 2014, we updated the PINT database⁷ to contain all published palaeointensity measurements from rocks older than 500 Myr at the site mean level and applied Q_{PI} criteria as set out in ref. 8 (see also <http://qpi.wikispaces.com/>). These are a set of 8 criteria based on the same model of Q criteria³¹ as widely used for palaeomagnetic poles but reformulated for palaeointensity measurements and applied at the site-mean level. The Q_{PI} value is the sum of the criteria met and is intended to comprehensively reflect the extent to which numerous known sources of bias to palaeointensity estimates have been reasonably guarded against.

An additional criterion 'MAG' was added to the original 8 criteria outlined in ref. 8, which stipulates that the associated raw measurement data must be publicly available for scrutiny (the MagIC database <http://earthref.org/MAGIC/> provides the ideal venue for these). It was recently observed that thermoremanent magnetization (TRM) preserved in non-single domain grains can be meta-stable, producing an additional potential source of bias to palaeointensity estimates³². A very recent study³³ has suggested that this bias could be towards either over- or underestimation of the palaeointensity (depending on the magnetic history of the samples) but could be guarded against by applying sufficiently strict reliability criteria. Importation of the raw data into the MagIC database would allow this to be done at a future date.

A further change to the criteria outlined in ref. 8 is that, in the present study, to meet the 'AGE' criterion, we required the maximum nominal uncertainty in the age estimate of the result to be less than or equal to 50 Myr.

For the purpose of testing the dipole relationship (Fig. 1 and Extended Data Fig. 6), we accepted only measurements of the palaeointensity that had associated directional data (also at the site mean level) derived from a minimum of three specimens with an associated Fisher precision parameter $k > 10$ and/or 95% cone of confidence $\alpha_{95} < 30^\circ$.

Significant correlations between palaeointensity and inclination were tested for, using the Kendall τ rank correlation coefficient assessed at the 5% significance level. The one-tailed correlation was used to test specifically for a positive rank correlation, which would be expected for a dominantly dipolar field.

Virtual axial dipole moments (VADMs) and virtual dipole moments (VDMs), collectively referred to here as VDMs, are calculated (in ZAm^2 , where $1 \text{ ZAm}^2 = 10^{21} \text{ Am}^2$) using equation (1)

$$\text{VDM} = \frac{4\pi r^3}{\mu_0} F(1 + 3\cos^2\theta)^{-\frac{1}{2}} \quad (1)$$

where r is the radius of the Earth ($6.371 \times 10^6 \text{ m}$), μ_0 is the permeability of free space ($1.257 \times 10^{-6} \text{ m kg s}^{-2} \text{ A}^{-2}$), F is the palaeointensity (in Tesla) and θ is the magnetic colatitude calculated using equation (2):

$$\theta = 90^\circ - \tan^{-1}(1/2 \tan I) \quad (2)$$

where I is the site mean inclination for a virtual dipole moment (that is, assuming a dipole field) and I is the study mean inclination for a VADM (that is, assuming an axial dipole field and a sufficient averaging of directional secular variation)

Monte Carlo resampling test. To test further the hypothesis that the 'Mid' period ($\sim 1,300$ – $2,400$ Myr ago) has VDMs that are significantly lower than those of the 'Early' ($> 2,400$ Myr ago) and 'Late' ($< 1,300$ Myr ago) periods we adopt a Monte Carlo resampling approach with 10,000 repetitions. For this, we consider the uncertainties in both the ages and VDMs and therefore exclude data where no uncertainties are reported, which leaves a total of 183 results with $Q_{PI} \geq 1$. For each repetition, the age and VDM of each result are resampled from normal distributions where the means are the reported mean values. The reported age uncertainties are taken to represent $2\sigma_{\text{age}}$ errors where σ_{age} is the standard deviation. For the VDM standard deviations we use the unbiased estimate of the standard deviation of the distribution of VDM means:

$$\sigma_{\text{VDM}} = t_{(1-\frac{0.95}{2}, N-1)} \frac{\text{VDM}_{\text{err}}}{\sqrt{N}} \quad (3)$$

where t is the t -critical value for the 68th percentile (that is, the standard deviation coverage interval for a normal distribution), VDM_{err} is the reported uncertainty, and N is the number of specimens used to estimate the mean. σ_{VDM} represents the VDM distribution that would be obtained if we were able to repeat the experiments multiple times. After all data have been resampled for a given repetition, the resampled data are split into the 'Early', 'Mid' and 'Late' periods and a one-tailed Kolmogorov–Smirnov test for equality of distributions is performed.

We count the proportion of repetitions where we cannot reject the null hypothesis of the Kolmogorov–Smirnov test. This represents the proportion of repetitions where it is unlikely that the 'Mid' period VDMs are lower than the other periods at the 5% significance level. The results for the test with various Q_{PI} thresholds are given in Extended Data Table 1. Despite the reduced number

of data, the resampling test, which accounts for data uncertainties, confirms a reduction in the average dipole moment of up to $Q_{PI} \geq 3$. However, too few data, particularly for the 'Mid' period, are available to confirm this for $Q_{PI} \geq 4$ – 5 .

New likelihood test. A potential problem with using general statistical tests to determine significance for palaeomagnetic data sets is that they do not account for potentially strong correlations that may occur between data that are typically sampled highly non-uniformly through time. Here we are attempting to isolate variations on the billion-year timescale but we need to consider the risk that observed differences in fact arise from over-sampling of periods of unusually high- or low-field geomagnetic intensity produced by shorter-timescale variations.

Specifically, either mantle convection may change the heat flowing across the core–mantle boundary, causing shifts in the dipole moment lasting tens of millions of years²⁶ (process 1), or secular variation, reflecting the intrinsic operation of the geodynamo, may produce similar shifts lasting up to a few hundred thousand years²¹ (process 2).

Our understanding of the above processes is incomplete even for recent times and is very poor for the Precambrian period with which we are concerned here. Nevertheless, we designed a test which attempted to incorporate process 2 by using a record of dipole moment variations for the last 2 million years²⁰ and, further, allowed this to be rescaled (producing variations in the long term average of up to a factor of three) to account for process 1. The test was later repeated, allowing for variations of up to a factor of 6 and a factor of 12 from process 1.

It is impossible to be certain whether our rescaled models are representative for the time periods being tested. Nevertheless, we point out that the minimum (5 ZAm^2), maximum (143 ZAm^2) and median (54 ZAm^2) values generated by our models (using a factor-of-three variation for process 1) do at least appear to be similar to those observed in the measured values that we are testing (Fig. 2). Also, the test outlined below compares relative rather than absolute differences, so a very good fit is not required.

For the purpose of the tailored likelihood test, we first assigned every measurement in our $Q_{PI} \geq 3$ subset to a 'mantle group' and a 'secular variation group' (Supplementary Table 2). Each mantle group comprised results with stated ages within 50 Myr of one another. Where estimates could be non-uniquely assigned to mantle groups, they were placed in with the estimates whose age was closest to their own. Each secular variation group comprised results that all had the same stated age. If results had the same age but were assigned different polarities, they were placed in separate secular variation groups.

The likelihood test estimated the probability of differences of the relative magnitude observed between the dipole moment distributions from two intervals ('Early', 'Mid' and/or 'Late') being arrived at by chance alone subject to the simulated effects of processes 1 and 2 above. First, for each real pair of data sets, the following were calculated: (1) the relative difference in the medians (expressed as a percentage of the smaller value) ($P(\text{Med})$ in Extended Data Table 2); (2) the relative difference in the interquartile ranges (IQRs; expressed as a percentage of the smaller value) ($P(\text{IQR})$ in Extended Data Table 2); and (3) the P -value associated with a Kolmogorov–Smirnov test for equality of distribution ($P(\text{K-S})$ in Extended Data Table 2).

Subsequently, these were compared to similar values produced by two pseudo-data sets which were of equal size to the real data sets and which contained identically sized and configured mantle groups and secular variation groups. These pairs of pseudo-data sets were derived by 10,000 iterations of the following procedure (see example in Extended Data Fig. 7):

(1) For each mantle group, the PADM2M model²⁰ of dipole moment variations for the last 2 Myr was rescaled using a factor drawn at random from a uniform distribution with a range of 0.5–1.5 (0.375–2.25 for rescaling factor 6 and 0.25–3.00 for rescaling factor 12). This was done to incorporate variations that might plausibly arise from mantle forcing of the geodynamo into the test.

(2) For each secular variation group within the mantle group, a 200-kyr continuous sub-interval within the rescaled model was selected at random. This was done to allow for data from the same secular variation group to be plausibly clustered in time.

(3) For each measurement within the secular variation group, a dipole moment estimate was randomly selected from within the sub-interval.

The likelihood of obtaining each of the three values in the likelihood test (above) by chance alone ($P(\text{Med})$, $P(\text{IQR})$ and $P(\text{K-S})$ in Extended Data Table 2) was estimated by the fraction of the 10,000 randomly generated pseudo-data sets that produced differences of the same or larger magnitude in these values.

The likelihood of simultaneously obtaining such differences ($P(\text{ALL})$ in Extended Data Table 2) was estimated by the fraction of the 10,000 randomly generated pseudo-data sets that produced differences of the same or larger magnitude in all three of the values ($P(\text{Med})$, $P(\text{IQR})$ and $P(\text{K-S})$ in Extended Data Table 2) simultaneously.

‘Recent’ data set. Our ‘Recent’ data set consists of all measurements of VDM in the PINT database at the site-mean level that are derived from rocks with a stated age between 1 Myr and 500 Myr. The interval 0–1 Myr was excluded to minimize skewing of the data set and estimates were further required to meet the following two criteria:

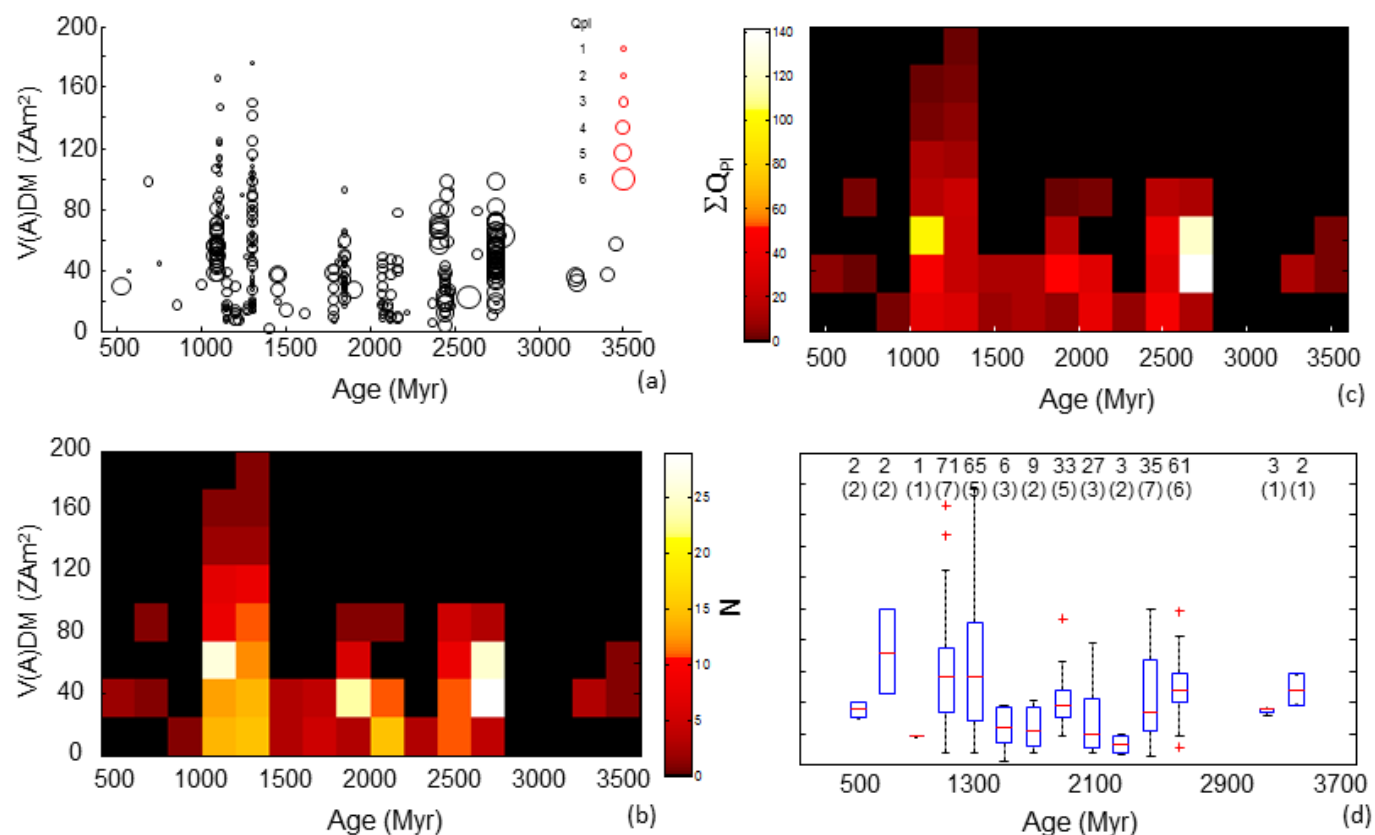
(1) The Q_{PI} criterion ‘STAT’ (ref. 8) which stipulates that the number of sample palaeointensity measurements comprising the mean is ≥ 5 and that the associated standard deviation is $\leq 25\%$ of the mean value.

(2) The use of one of the following palaeointensity techniques: T+ (Thellier with pTRM checks), M+ (Microwave with pTRM checks), LTD-DHT Shaw or some combination of techniques including at least one of the above. This should ensure that all results meet the ‘ALT’ criterion in ref. 8 of the Q_{PI} set.

These criteria were chosen as they are two of the most important indications of reliability and can easily be applied to measurements in the PINT database without

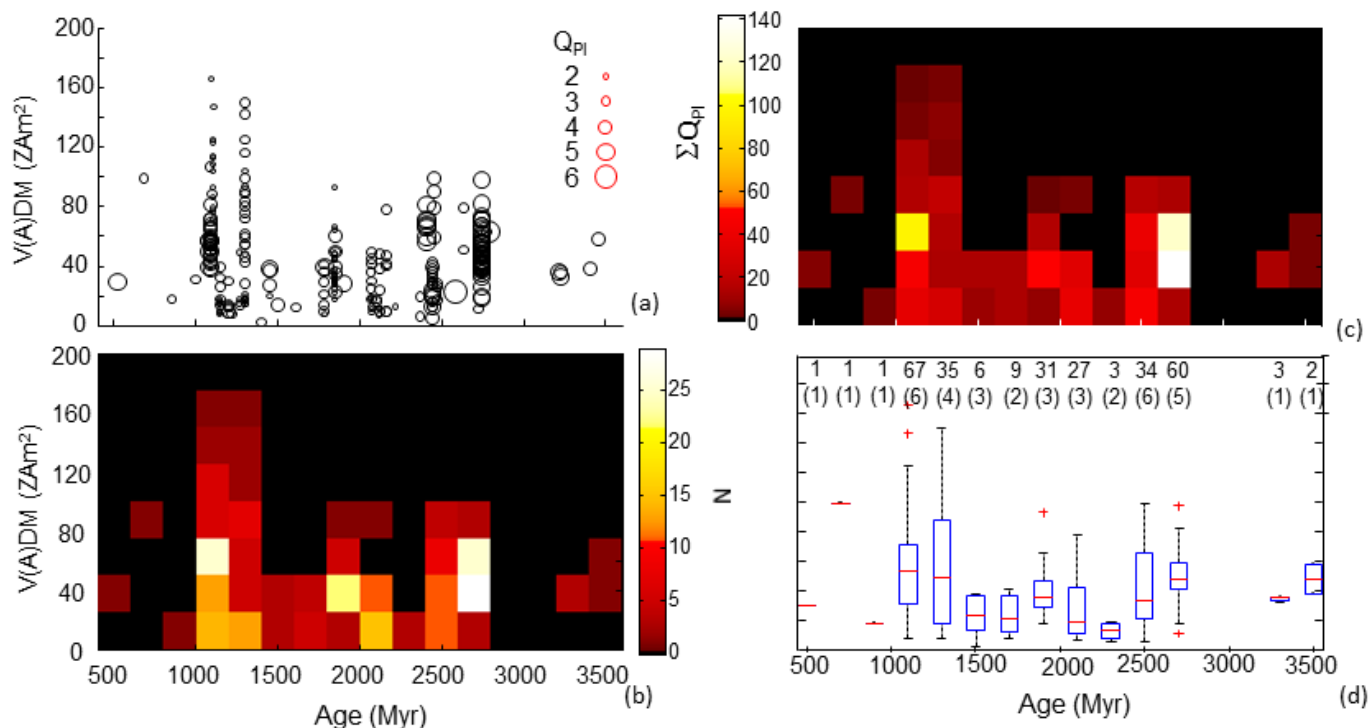
requiring the original manuscript to be rechecked. They should ensure that the ‘Recent’ data set comprises measurements with associated Q_{PI} values of at least 2. Previous experience suggests that the vast majority of this data set will also satisfy the ‘AGE’ criterion in ref. 8 and many will also satisfy others too. We therefore expect that the median Q_{PI} for the 218 estimates comprising the ‘Recent’ data set to be either 3 or 4. Some 139 estimates had directional information meeting the requirements defined above and the palaeointensity–inclination relationship is plotted in Extended Data Fig. 6b.

31. Van der Voo, R. The reliability of paleomagnetic data. *Tectonophysics* **184**, 1–9 (1990).
32. Shaar, R. & Tauxe, L. Instability of thermoremanence and the problem of estimating the ancient geomagnetic field strength from non-single-domain recorders. *Proc. Natl Acad. Sci. USA* **112**, 36, <http://dx.doi.org/10.1073/pnas.1507986112> (2015).



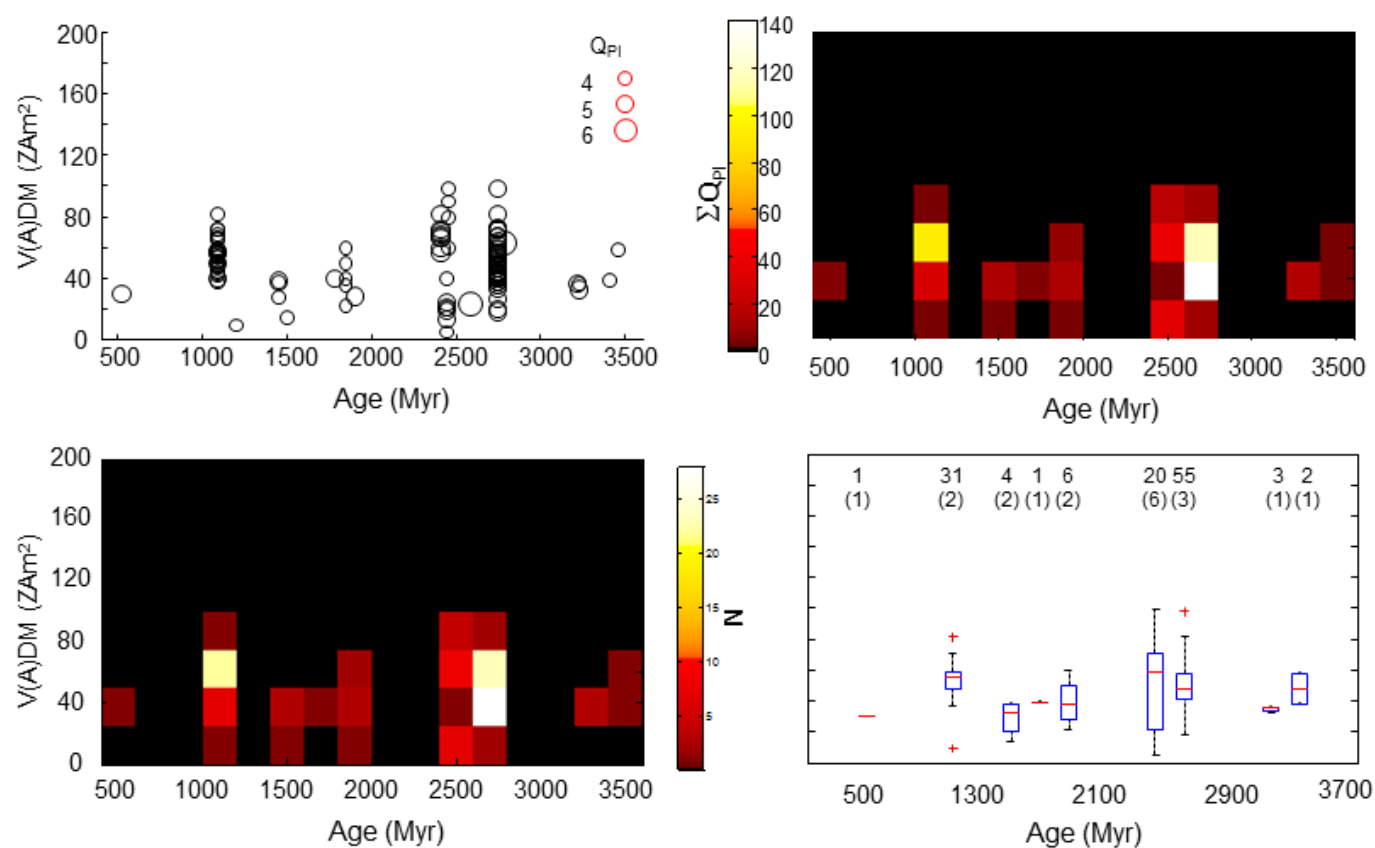
Extended Data Figure 1 | Four different representations of VDM versus time for all data with $Q_{PI} \geq 1$. **a**, Bubble plot, where size indicates Q_{PI} value. **b**, Density plot of number of measurements. **c**, Density plot of sum of Q_{PI}

values. **d**, Box plot after binning with an interval length of 200 Myr (number of data in each are given with the number of published studies in parentheses). See Fig. 1 caption for an explanation of the box plot.



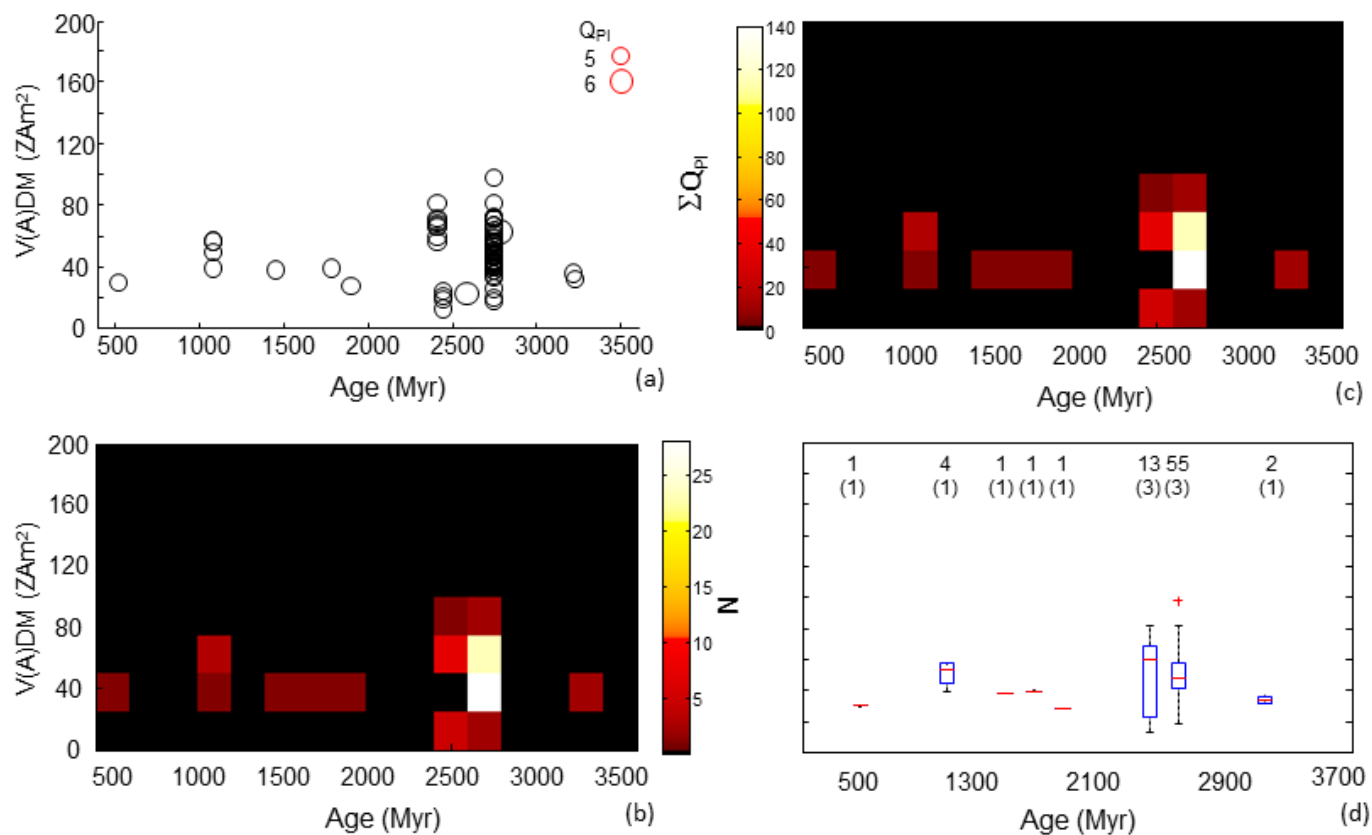
Extended Data Figure 2 | Four different representations of VDM versus time for all data with $Q_{PI} \geq 2$. **a**, Bubble plot, where size indicates Q_{PI} value. **b**, Density plot of number of measurements. **c**, Density plot of sum of Q_{PI}

values. **d**, Box plot after binning with an interval length of 200 Myr (number of data in each are given with the number of published studies in parentheses). See Fig. 1 caption for an explanation of the box plot.



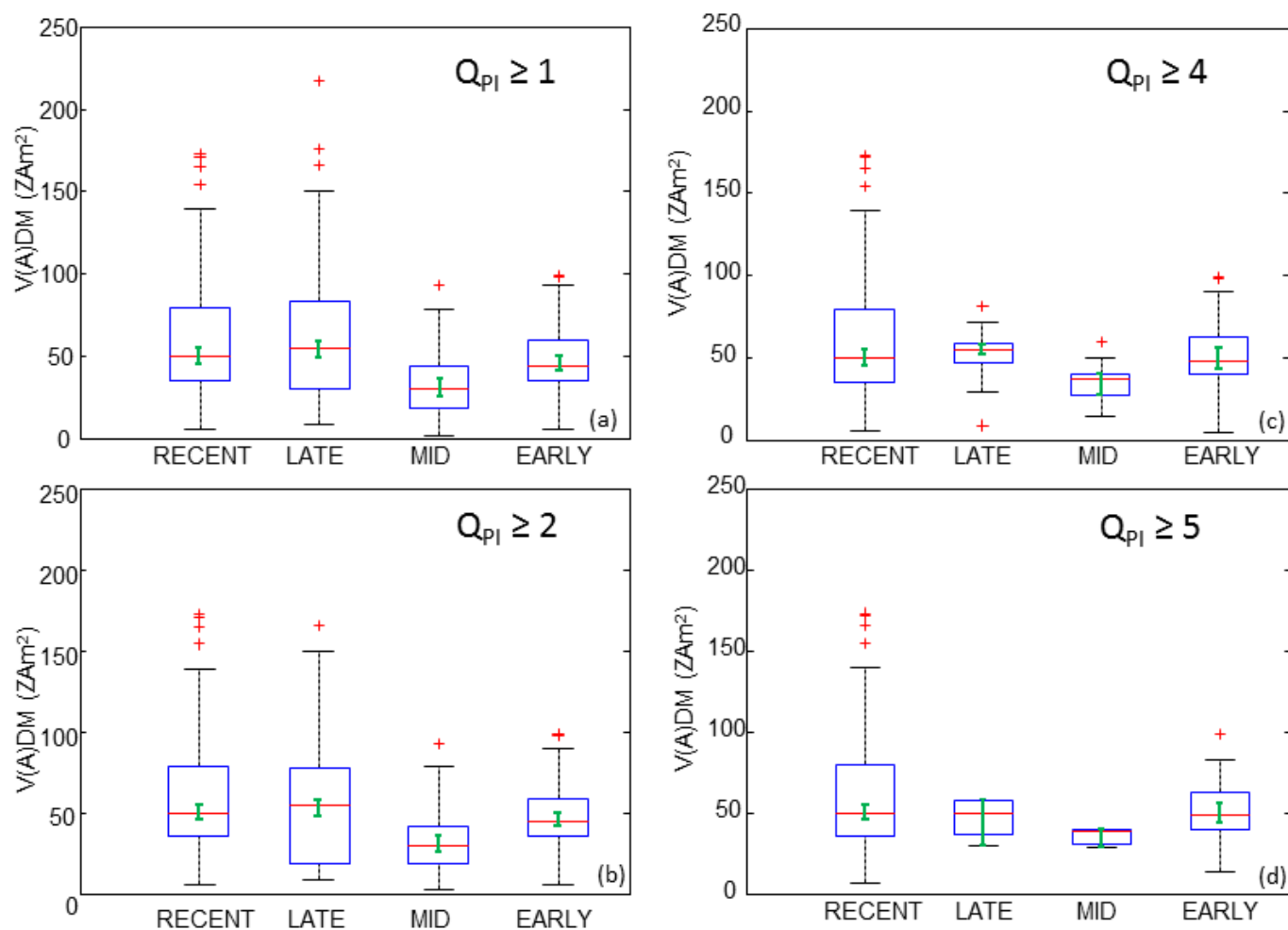
Extended Data Figure 3 | Four different representations of VDM versus time for all data with $Q_{PI} \geq 4$. **a**, Bubble plot, where size indicates Q_{PI} value. **b**, Density plot of number of measurements. **c**, Density plot of sum of Q_{PI}

values. **d**, Box plot after binning with an interval length of 200 Myr (number of data in each are given with the number of published studies in parentheses). See Fig. 1 caption for an explanation of the box plot.



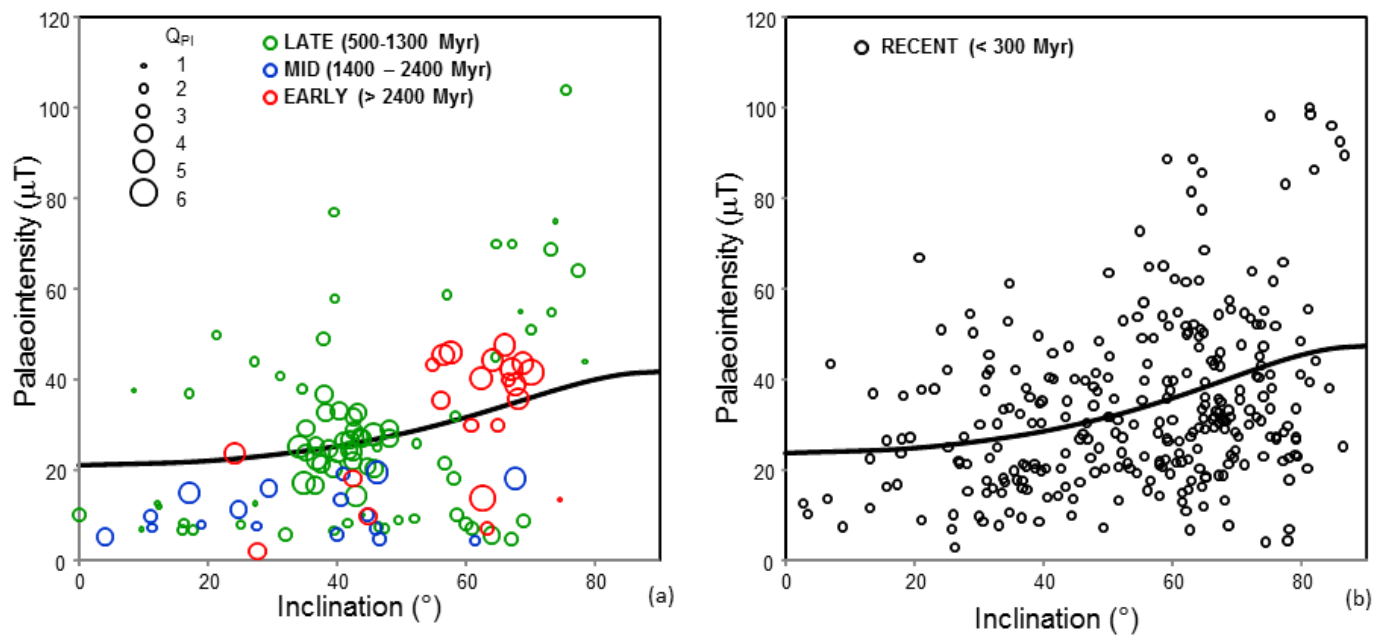
Extended Data Figure 4 | Four different representations of VDM versus time for all data with $Q_{PI} \geq 5$. **a**, Bubble plot, where size indicates Q_{PI} value. **b**, Density plot of number of measurements. **c**, Density plot of sum of Q_{PI}

values. **d**, Box plot after binning with an interval length of 200 Myr (number of data in each are given with the number of published studies in parentheses). See Fig. 1 legend for an explanation of the box plot.

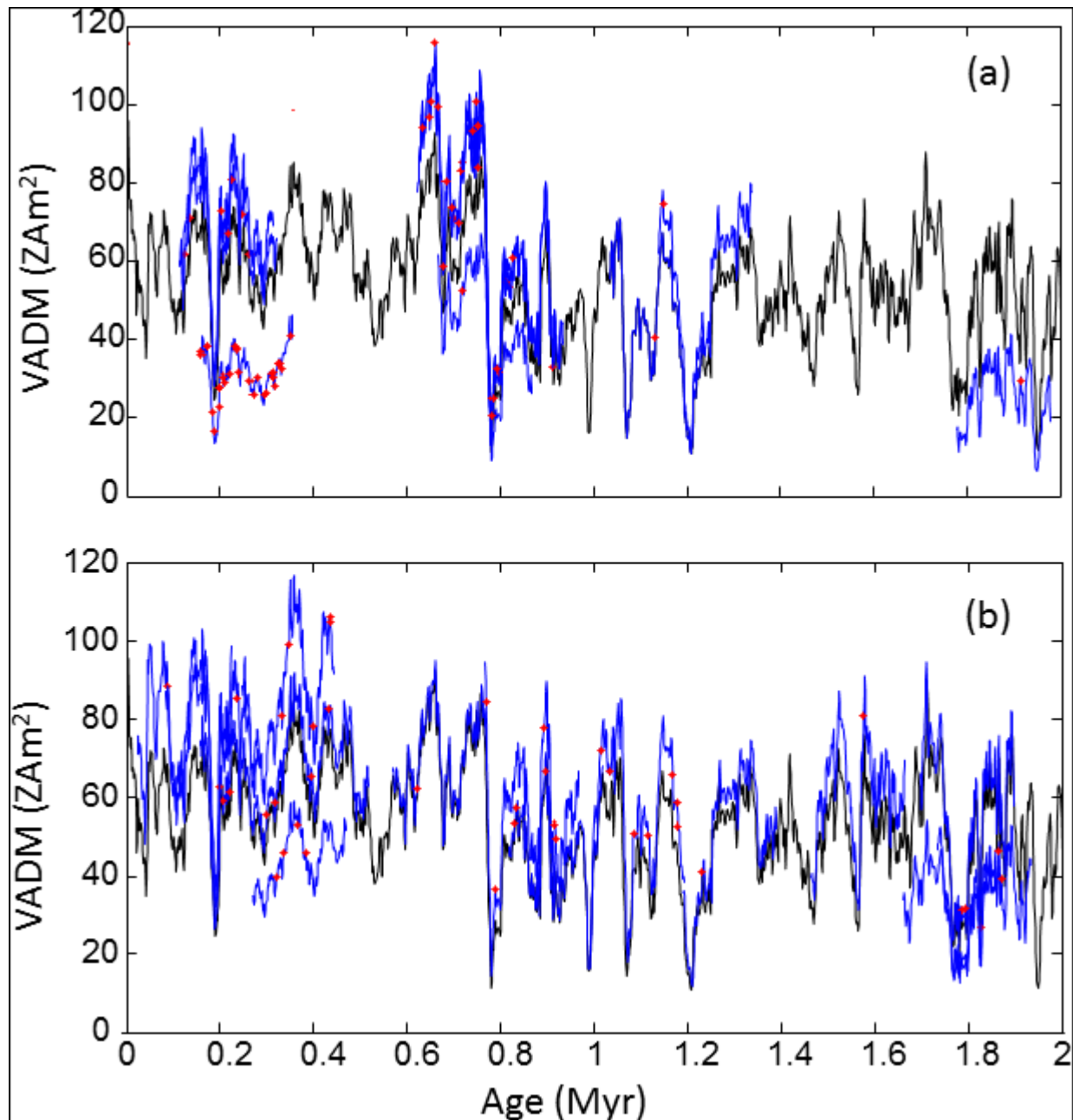


Extended Data Figure 5 | Box-plots for time intervals defined in the main text and summarized in Supplementary Table 3, comprising measurements with different minimum Q_{pl} values. See Fig. 3 for $Q_{pl} \geq 3$ plot and Fig. 1

caption for an explanation of the box plot. Thick error bars indicate 95% confidence limits (from 10,000 bootstraps) on the medians.



Extended Data Figure 6 | Raw palaeointensity versus inclination data shown with a best-fitting dipole for the four studied time intervals. Circle size indicates Q_{PI} value in panel a. In each case, the best-fitting dipole was found using the least-squares approach.



Extended Data Figure 7 | Examples of two pseudo-data sets produced by one iteration of the new likelihood test. See Methods for details. Each of the VADM estimates (red asterisks) are drawn from one of multiple 200-kyr-long sub-intervals (blue lines) of PADM2M²⁰ (black line) which is rescaled by a random factor between 0.5 and 1.5. Data from the same mantle group are drawn from sub-intervals with the same rescaling to simulate the possible

effects of mantle-forced variations. Data from the same secular variation groups are drawn from the same 200-kyr sub-interval to simulate the possible effects of further temporal clustering. Panel **a** shows an example using the mantle groups and secular variation groups of interval 'Late' and panel **b** shows the same for interval 'Mid'.

Extended Data Table 1 | Summary results from the Monte Carlo resampling test

| | | | | % of MC repetitions where MID V(A)DMs are unlikely to be less than other periods | | |
|---------------------|--------------------|------------------|-------------------|--|-------|--------------|
| | N _{EARLY} | N _{MID} | N _{LATE} | EARLY | LATE | EARLY + LATE |
| Q _{pI} ≥ 1 | 72 | 47 | 64 | 0.0 | 0.0 | 0.0 |
| Q _{pI} ≥ 2 | 71 | 47 | 64 | 0.0 | 0.0 | 0.0 |
| Q _{pI} ≥ 3 | 71 | 24 | 46 | 0.0 | 0.2 | 0.0 |
| Q _{pI} ≥ 4 | 62 | 7 | 32 | 95.2 | 45.0 | 83.1 |
| Q _{pI} ≥ 5 | 51 | 1 | 5 | 100.0 | 100.0 | 100.0 |

See Methods for details. *N* refers to the number of data (required to have quoted uncertainty values) used in each test.

Extended Data Table 2 | Results of the tailored likelihood test applied to data sets in Supplementary Table 2

| $Q_{pi} \geq 3$ | LATE vs MID | LATE vs MID (excl. Ref. 27) | LATE vs MID (excl. Ref. 28) | LATE vs MID (Factor 6) | LATE vs MID (Factor 12) | MID VS EARLY | LATE VS EARLY |
|-----------------|--------------|--------------------------------|--------------------------------|---------------------------|----------------------------|--------------|---------------|
| P(Med) | 0.029 | 0.123 | 0.089 | 0.174 | 0.267 | 0.202 | 0.563 |
| P(IQR) | 0.543 | 0.021 | 0.800 | 0.607 | 0.636 | 1.000 | 0.579 |
| P(K-S) | 0.060 | 0.222 | 0.223 | 0.107 | 0.129 | 0.342 | 0.843 |
| P(ALL) | 0.015 | 0.006 | 0.068 | 0.061 | 0.084 | 0.187 | 0.348 |

See Methods for details. $P(\text{Med})$ and $P(\text{IQR})$ refer to the estimated likelihoods of the observed differences in the medians and interquartile ranges arising by chance alone. $P(\text{K-S})$ is the same but refers to the given level of significance observed in the Kolmogorov–Smirnov test of equality for probability distributions. $P(\text{ALL})$ refers to likelihood of all three above likelihoods being met simultaneously by chance alone. Note that the ‘Late’ versus ‘Mid’ tests were repeated after excluding data from studies referred to in the main text in order to test the robustness of the observed differences.

Sex-specific demography and generalization of the Trivers–Willard theory

Susanne Schindler¹, Jean-Michel Gaillard², André Grüning³, Peter Neuhaus⁴, Lochran W. Traill⁵, Shripad Tuljapurkar⁶ & Tim Coulson¹

The Trivers–Willard theory¹ proposes that the sex ratio of offspring should vary with maternal condition when it has sex-specific influences on offspring fitness. In particular, mothers in good condition in polygynous and dimorphic species are predicted to produce an excess of sons, whereas mothers in poor condition should do the opposite. Despite the elegance of the theory, support for it has been limited^{2,3}. Here we extend and generalize the Trivers–Willard theory to explain the disparity between predictions and observations of offspring sex ratio. In polygynous species, males typically have higher mortality rates⁴, different age-specific reproductive schedules and more risk-prone life history tactics than females; however, these differences are not currently incorporated into the Trivers–Willard theory. Using two-sex models parameterized with data from free-living mammal populations with contrasting levels of sex differences in demography, we demonstrate how sex differences in life history traits over the entire lifespan can lead to a wide range of sex allocation tactics, and show that correlations between maternal condition and offspring sex ratio alone are insufficient to conclude that mothers adaptively adjust offspring sex ratio.

Trivers and Willard¹ proposed that when the fitness benefit to a mother of producing sons increases faster with her own condition than the benefit of producing a daughter, good-condition mothers should produce more male than female offspring. Trivers and Willard hypothesized that these fitness benefits should be observed when (1) maternal condition determines offspring condition, (2) the condition of offspring at independence correlates with condition at adulthood, (3) good-condition males produce more offspring than poor-condition ones, and (4) there is greater variation in lifetime reproductive success among males than females^{1,2}. The original theory focused on monotocous species with non-overlapping generations, but has been extended to polytocous populations⁵, and to overlapping generations⁶. The theory has been investigated in a wide range of bird⁷ and mammal³ species, including humans⁸, but has received mixed support^{2,9}, with many species expected to adhere to Trivers and Willard's predictions often failing to do so^{2,9}. The discrepancy between theoretical predictions and empirical observations has led to much debate, with explanations falling into three broad categories: first, it may be physiologically impossible for a female to determine the sex of her offspring⁵; second, tests of the theory are often inadequate as systems have not been shown to conform to the assumptions of the theory^{9,10}; and third, data quality is poor³. The consensus seems to be that adaptive sex ratio production of offspring often does not occur when expected^{2,3}. Modelling work has been performed to examine why this is the case. In a notable article, Leimar¹¹ demonstrated that appropriate tests of the Trivers–Willard theory¹ require a comparison of the reproductive values (RV) of sons and daughters rather than a comparison of lifetime reproductive success, as reproductive value is the appropriate measure of fitness¹¹. The lifetime reproductive success of an individual

measures the number of offspring produced over its lifetime, whereas RV describes the fraction of a future population that has descended from it¹². Leimar¹¹ went on to show that if maternal condition influences the RV of offspring in one sex more than in the other, then life histories in which good-condition mothers produce an excess of daughters could be adaptive. However, to our knowledge, no one has calculated the RV of females and males for naturally occurring systems. Leimar's model¹¹ consequently remains an elegant, but abstract, demonstration of theoretical scenarios where species, which otherwise conform to Trivers and Willard's assumptions, do not follow the predicted sex allocation tactic. Here we extend and generalize the insights of Leimar¹¹ and demonstrate that, contrary to Fisher's theory of parental investment¹³, sex differences in life history at both pre- and post-independence determine optimal offspring sex ratio as a function of maternal condition, and apply this new approach to two empirical data sets.

Where adaptive sex ratio variation has been expected, for instance in polygynous and sexually dimorphic species, males often have higher mortality rates than females at all stages of life^{4,14} as a consequence of their more risk-prone life history tactic which, in many species, involves reproducing at a later age and fighting more with conspecifics than females^{15,16}. Fisher's theory¹³ of sex allocation predicts higher investment in the rarer sex, or the sex experiencing higher mortality during the period of dependency, such that the sex ratio at independence is unity. However, this conclusion holds at the population level only under the assumption that all same-sex individuals have equal chances of reproducing. Trivers and Willard¹ refined Fisher's theory¹³ by demonstrating that, for a given individual, it might be optimal to invest more into the sex with higher fitness benefits if these vary with maternal condition. Remarkably, Trivers and Willard did not explicitly list sex differences in mortality or growth between independence and adulthood, or age differences in reproductive output as an adult, when identifying conditions in which their predictions should be observed. Sex differences in demographic rates have never been incorporated into models predicting the RV of female and male offspring of mothers in varying conditions^{6,11,17}. Although Fisher¹³ stated that the period after offspring independence should not influence the optimal sex ratio, the RV factors in information on mortality and reproductive rates at all ages. This means that when examining the relative RV of sons and daughters to a mother of a given condition, it is necessary to consider offspring mortality rates both pre- and post-independence, and reproductive rates at all ages.

Estimating RV from two-sex models is challenging as standard approaches to estimate RV for one-sex models do not easily extend to two sexes. Therefore, we first developed an approach to calculate RV for realistic two-sex models (see Methods). Next, we incorporated sex-specific mortality and other sex-specific life history traits into two-sex integral projection models¹⁸ (IPMs) and explored how the RV of male and female offspring born to mothers of a given condition vary as the

¹University of Oxford, Department of Zoology, Oxford OX1 3PS, UK. ²Laboratoire de Biométrie et Biologie Evolutive (UMR 5558), Université Claude Bernard Lyon 1, 43 boulevard du 11 novembre 1918, 69622 Villeurbanne Cedex, France. ³Department of Computer Science, University of Surrey, Guildford GU2 7XH, UK. ⁴Department of Biological Sciences, University of Calgary, Calgary, Alberta T2N 1N4, Canada. ⁵School of Animal, Plant and Environmental Sciences, University of the Witwatersrand, Wits 2050, South Africa. ⁶Department of Biology, Stanford University, Stanford, California 94305, USA.

level of sex differences in a life history trait is altered (see Methods). Specifically, we took a published model for Columbian ground squirrels (*Urocitellus columbianus*)¹⁸, a system where we neither expect nor observe females skewing offspring sex ratio as a function of body weight¹⁹, and examined under what circumstances we could generate optimal offspring sex ratios varying with maternal weight. We next adapted a published two-sex IPM for bighorn sheep (*Ovis canadensis*)²⁰, a species where adaptive sex ratio variation is expected but empirical evidence is inconclusive^{21–23}, and predicted optimal offspring sex allocation.

Using an IPM to explore adaptive sex ratio variation allowed us to alter a range of demographic rates in ways that are not possible in naturally occurring systems. For instance, we altered the strength of size selection in a mating system, differential male and female mortality schedules, assumptions about correlations between weight at independence and at maturity, and between parental and offspring weight. We calculated the expected RV of male and female offspring born to mothers of a given weight (a surrogate for condition widely used in empirical tests of the Trivers–Willard theory^{2,3,7,19,21,24,25}; see Methods). We then calculated the difference between the RV of male and female offspring at each maternal weight and the slope of this difference in relation to maternal weight (Fig. 1; Methods). A positive slope represents support for the prediction of Trivers and Willard¹ that large mothers should produce an excess of sons and small mothers an excess of daughters; we call this a ‘Trivers–Willard effect’ (Fig. 1a). A negative slope represents a reverse sex allocation tactic that we call a ‘reversed Trivers–Willard effect’ because large females gain a fitness advantage by producing daughters rather than sons (Fig. 1d). A slope not differing from zero indicates that offspring of both sexes provide equivalent fitness benefits to all mothers. We also identified two further sex allocation tactics in which the difference between male and female RV is u-shaped or n-shaped. In the first of these tactics (a ‘Trivers–Willard effect type 3’), mothers in poor or exceptional condition should produce sons, while all others should produce daughters (Fig. 1e). Likewise, we called the tactic in which only the lightest and heaviest mothers should produce daughters the ‘Trivers–Willard effect type 4’ (Fig. 1f). Next, we altered parameters in the model that determine the level of sex differences in life history traits and recalculated the slope. The difference between the slopes from the perturbed and the unperturbed model represents the sensitivity of adaptive sex ratio variation to altering aspects of the survival, fertility, growth, inheritance and mating functions that constitute the IPM. A positive value in the difference between the slopes implies an increase in the strength of the Trivers–Willard effect (Fig. 1b). A negative slope difference suggests a decrease in the strength of the Trivers–Willard effect (Fig. 1c), leading to a possible reversal of the prediction if the difference is sufficiently negative (Fig. 1d). A parameter change can also cause the difference between male and female RV to be u-shaped or n-shaped (Fig. 1e, f), which might apply to systems with two alternative reproductive morphs (‘sneakers’ and ‘fighters’). Our framework allowed us to assess how changing the level of sex differences in life history tactics, and consequently the level of sex differences in demographic rates, affects the direction of selection on sex ratio variation for a given maternal condition.

For this, we used a recently published two-sex IPM of Columbian ground squirrels¹⁸—a species that is polytocus²⁶, polygynandrous²⁷, and has greater variance in reproductive success among males than females²⁶. Differences in male and female mortality schedules are less pronounced than in most species where Trivers–Willard effects are expected²⁸. We therefore treated their demographic rates as identical (Supplementary Table 1). We predicted a small reversed Trivers–Willard effect: a small deviation from a sex ratio of unity as maternal weight increases (Fig. 2; solid black line). The difference in RV is small because we assumed identical survival and growth rates for females and males. The reason a reverse Trivers–Willard effect occurs is that, although males and females have identical growth and survival rates in our model, males need to reach a greater age before achieving good

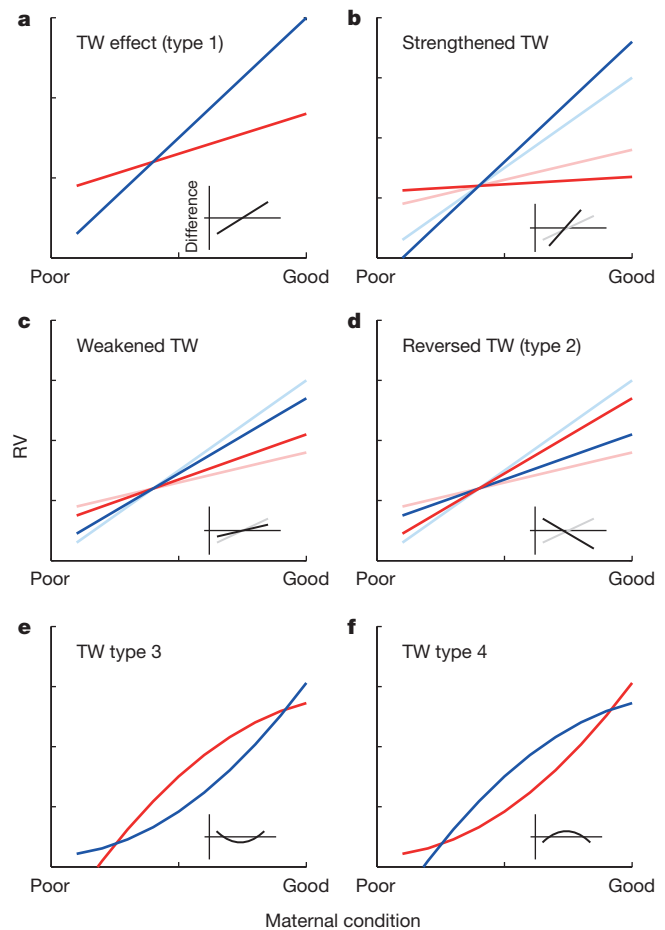


Figure 1 | Female reproductive value (RV, red line) and male RV (blue) depend on maternal condition. Differences between male and female RV are shown as insets (black and grey lines). **a**, Trivers–Willard (TW) effect: for good-condition mothers, sons have higher RV than daughters. For poor-condition mothers, daughters have higher RV than sons. **b–d**, Changes in demographic parameters can result in a more pronounced Trivers–Willard effect (**b**), a smaller difference between female and male RV (**c**), or a reversal of a Trivers–Willard effect (**d**). Original RV and difference in RV before demographic rates have been changed are shown in light red (females), light blue (males) and grey (difference in RV in insets) in **b–d**. **e, f**, The difference between male and female RV can be u-shaped (type 3 Trivers–Willard effect) (**e**), or n-shaped (type 4 Trivers–Willard effect) (**f**).

chances of reproductive success compared to females because male mating success is size-selective. Large females maximize their RV, when litter size is fixed, by producing females that breed at a younger age instead of producing males. We can see this by altering the mean age of reproduction among females; as this increases, the reversed Trivers–Willard effect is reduced until it eventually disappears completely (Extended Data Fig. 1a, b).

When we increased mortality of pre-reproductive males (that is, squirrels below an estimated size threshold of 279 g) to be 3% greater than female mortality, we predict that the reversed effect is replaced with a Trivers–Willard effect (Fig. 2; dotted line). As male mortality increases, fewer males make it to reproductive age, but those that do (that is, those born to larger mothers) have increased reproductive success. Mothers that can produce large males with a reasonable chance of growing to reproductive age can gain substantial RV because males are rare, even though males may have to wait longer to achieve reproductive success. In other words, as we alter differential mortality between the sexes and the variance in reproductive success amongst males, we alter the optimal tactic that mothers of a given body size should follow (Extended Data Fig. 2, solid line; see Extended Data Fig. 3 for a trade-off between mortality and mating chances).

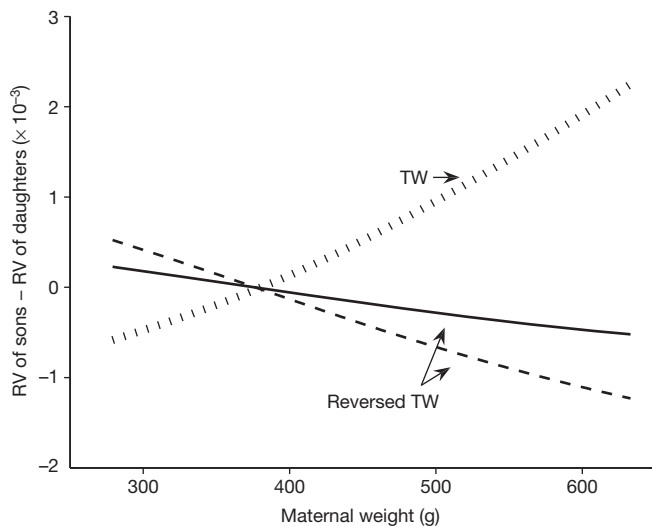


Figure 2 | Reproductive value differs with maternal condition and offspring sex in Columbian ground squirrels. The graph shows RV of sons minus RV of daughters when male mating success increases with body weight. Solid line, females and males have the same mortality rate; dotted line, survival rates of pre-breeding males (weight below 279 g) are 3% lower than female survival rates; dashed line, increased size selection in male mating success (parameter ρ raised from 0.1 to 0.25, see Supplementary Table 1). We scaled the RV of females and males such that the female RV of the smallest reproductive size class is 1.

However, these results are context-dependent. In our initial model, the population growth rate was greater than one and the population increased in size with time. If we increased female mortality such that the population growth rate fell below one, the advantage of reproducing early disappeared, and females could gain reproductive value by producing males that mate at a later age compared to females.

These results show that because RV is a complex function of condition-dependent mortality and reproductive rates, and these rates generate a growing or shrinking population, the optimal sex allocation tactic can change over a relatively small range of parameter values. Despite this, it is valuable to analyse our baseline model parameterized for squirrels. As parameter values are altered, we find that those parameters generating a differential between male and female mortality (survival function), and those that influence the correlation between weight at independence and weight at adulthood (the growth function) strongly influence the expected bias in offspring sex ratio for a given maternal condition (Extended Data Figs 4 and 5). In contrast, altering age-independent female reproductive success (fecundity function) and the correlation between mean parental weight at time t and offspring mass when they recruit to the population at $t + 1$ (the inheritance function) have relatively small effects on the direction or magnitude of the Trivers–Willard effect (Extended Data Figs 4 and 5). However, the results are surprisingly nuanced. We find that increasing male mortality at all ages on the logit-scale by reducing the survival intercept generates a type 3 Trivers–Willard effect. In contrast, if we reduce the body-mass survival slope for males such that larger males have rates of mortality that are elevated to a greater extent than those that are smaller, we weaken the strength of the reversed Trivers–Willard effect. Despite this complexity, we identified a general pattern that holds for the first two types of sex allocation tactics (Trivers–Willard effect and reversed Trivers–Willard effect): if a perturbation to a parameter increases the variance in male RV at birth relative to the variance in female RV at birth, the strength of the predicted Trivers–Willard effect increases, and vice versa (Extended Data Fig. 5). This pattern holds for the bighorn sheep model.

Bighorn sheep are a monotonous, polygynous species with strong sexual size-dimorphism, and are expected to exhibit a Trivers–Willard

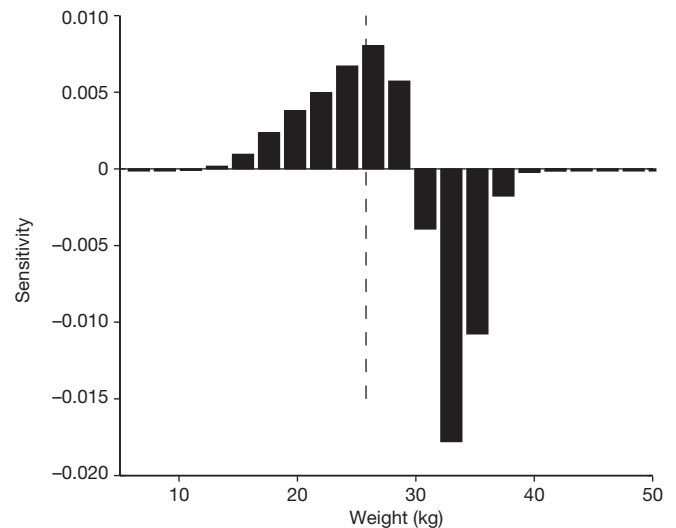


Figure 3 | Size-specific sensitivity of the Trivers–Willard effect to a 1% increase in male mortality in bighorn sheep lambs. The probability of mortality has been independently increased by 1% in each size class and the change of the slope of the Trivers–Willard effect plotted. The vertical dashed line denotes mean male lamb weight. Increasing mortality for male lambs below average weight strengthens the Trivers–Willard effect (bars are above zero), whereas increasing mortality for male lambs of above average weight weakens the Trivers–Willard effect (bars are predominantly below zero). Size-specific sensitivities of all stages (lamb, yearling, adult, and senescent) are provided in Extended Data Fig. 8. The y-axis plots $\partial \Delta v_a / \partial s_i \times s_i \varepsilon$, where $\partial \Delta v_a / \partial s_i$ is the sensitivity of difference between male and female reproductive value (Δv_a) to a perturbation in parameter s_i by ε (see Methods). The index a denotes age, shown here at birth (that is, $a = 0$); i denotes the parameter, running from 1–69 for the bighorn sheep model.

effect; empirically, however, they do not^{21,22}. In contrast to Columbian ground squirrels, female and male mortality schedules differ markedly throughout life¹⁴. We used a two-sex IPM including age structure²⁰ (see parameters in Supplementary Tables 2 and 3) because demographic rates depend on both weight and age in bighorn sheep²⁹. For this species, our model predicts a Trivers–Willard effect (Extended Data Fig. 6). The overall size of the Trivers–Willard effect in sheep is larger than for squirrels because there are marked sex differences in life history at all ages. However, as with squirrels, sex differences in mortality rates before age at first breeding had the same effects on the predicted Trivers–Willard effect: increasing mortality rates of male lambs and yearlings increased the strength of the Trivers–Willard effect, and increasing male mortality and male growth rates had the largest effects on the sex allocation tactic (Extended Data Fig. 7). In addition, within each life stage, increasing mortality rates of small males increased the strength of the Trivers–Willard effect, whereas increasing mortality of large rams had the opposite effect (Fig. 3 and Extended Data Fig. 8). This reveals that sex differences in size-specific mortality can trade-off against one another at different ages to influence the optimal behaviour of a mother of a given size. In spite of the similarities between squirrels and sheep, there are also differences. In sheep, parameters of the inheritance functions (which determine the correlation between parental and offspring size) also influence the magnitude of the effect (Extended Data Fig. 7). Taken together, these results reveal two things: first, there is a need to consider the entire life history when assessing whether a mother in a specific condition should produce male or female offspring; and second, sex differences in life history traits, especially before the age at first breeding and often beyond the age of dependency, can affect the RV of male and female offspring of a mother in a given condition.

Our results show that in order to conclude that females are behaving adaptively from a regression of maternal condition against offspring sex ratio, it is necessary to show (1) that offspring sex ratio is equivalent

to the ratio of male to female RV at birth, and (2) that the slope or shape of the association between maternal condition and difference in the RV of sons and daughters (Fig. 1) is of the same sign or form as the slope or shape of the regression of maternal condition against birth sex ratio. Offspring sex ratio alone may not be a good predictor of whether females are behaving adaptively because it does not necessarily strongly correlate with the relative RV of sons and daughters. This is because the RV is a complex function of age- and condition-specific survival, fertility, development trajectories and phenotypic inheritance. The calculation of RV for females and males is challenging and we have been unable to identify shortcuts to its calculation. Without this, it will be necessary to construct two-sex models of the population dynamics of species of interest and calculate the RV of sons and daughters born to mothers in different conditions. We are hopeful that application of our approach to multiple systems will reveal a shortcut to RV calculation. Until that time, we cannot conclude that females are, or are not, behaving adaptively by correlating offspring sex ratio with maternal condition alone.

We have extended Trivers and Willard's original theory¹ to incorporate differential demographic rates between the sexes. We show that currently the only way to examine how sex differences in mortality in a specific system influence optimal sex allocation tactic as a function of maternal condition is to calculate the RV for male and female offspring born to mothers in different conditions, and we have provided a method for doing so. As further research utilizing data to construct two-sex IPMs for different systems is carried out, we expect to gain a better understanding of the frequency with which females adaptively adjust the sex ratio of their offspring. As it is currently difficult to interpret the vast empirical literature testing this theory, calculating the RV of males and females born to mothers in different conditions for many systems will reveal whether Trivers and Willard's important insight that mothers should manipulate the sex ratio of their offspring to maximise their fitness is supported or not. However, the context-dependent nature of relative RV of sons and daughters to a mother suggests that an appropriate empirical test of the Trivers–Willard theory would be to observe mothers altering their sex allocation tactic across a range of conditions that affect the optimal sex allocation tactic.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 12 August 2014; accepted 22 July 2015.

Published online 21 September 2015.

1. Trivers, R. L. & Willard, D. E. Natural selection of parental ability to vary the sex ratio of offspring. *Science* **179**, 90–92 (1973).
2. Hewison, A. J. M. & Gaillard, J.-M. Successful sons or advantaged daughters? The Trivers–Willard model and sex-biased maternal investment in ungulates. *Trends Ecol. Evol.* **14**, 229–234 (1999).
3. Sheldon, B. C. & West, S. A. Maternal dominance, maternal condition, and offspring sex ratio in ungulate mammals. *Am. Nat.* **163**, 40–54 (2004).
4. Clutton-Brock, T. H. & Isvaran, K. Sex differences in ageing in natural populations of vertebrates. *Proc. Biol. Sci.* **274**, 3097–3104 (2007).
5. Williams, G. C. Question of adaptive sex-ratio in outcrossed vertebrates. *Proc. Biol. Sci.* **205**, 567–580 (1979).
6. Schwanz, L. E., Bragg, J. G. & Charnov, E. L. Maternal condition and facultative sex ratios in populations with overlapping generations. *Am. Nat.* **168**, 521–530 (2006).

7. Wiebe, K. L. & Bortolotti, G. R. Facultative sex ratio manipulation in American kestrels. *Behav. Ecol. Sociobiol.* **30**, 379–386 (1992).
8. Ruckstuhl, K. E., Colijn, G. P., Amiot, V. & Vinish, E. Mother's occupation and sex ratio at birth. *BMC Public Health* **10**, 269 (2010).
9. Brown, G. R. Sex-biased investment in nonhuman primates: can Trivers & Willard's theory be tested? *Anim. Behav.* **61**, 683–694 (2001).
10. Cockburn, A., Legge, S. & Double, M. in *Sex Ratios: Concepts and Research Methods*, (ed. Hardy, I. C. W.) Ch. 13, 266–286 (Cambridge Univ. Press, 2002).
11. Leimar, O. Life-history analysis of the Trivers and Willard sex-ratio problem. *Behav. Ecol.* **7**, 316–325 (1996).
12. Taylor, P. D. Allele-frequency change in a class-structured population. *Am. Nat.* **135**, 95–106 (1990).
13. Fisher, R. *The Genetical Theory of Natural Selection* (Oxford University Press, 1930).
14. Loison, A., Festa-Bianchet, M., Gaillard, J.-M., Jorgenson, J. & Jullien, J. Age-specific survival in five populations of ungulates: evidence of senescence. *Ecology* **80**, 2539–2554 (1999).
15. Lawson Handley, L. & Perrin, N. Advances in our understanding of mammalian sex-biased dispersal. *Mol. Ecol.* **16**, 1559–1578 (2007).
16. Kappeler, P. *Verhaltensbiologie* (Springer, 2006).
17. Charnov, E. L. *The Theory of Sex Allocation* (Princeton Univ. Press, 1982).
18. Schindler, S., Neuhaus, P., Gaillard, J.-M. & Coulson, T. The influence of nonrandom mating on population growth. *Am. Nat.* **182**, 28–41 (2013).
19. Gedir, J. V. & Michener, G. R. Litter sex ratios in Richardsons ground squirrels: long-term data support random sex allocation and homeostasis. *Oecologia* **174**, 1225–1239 (2014).
20. Traill, L. W., Schindler, S. & Coulson, T. Demography, not inheritance, drives phenotypic change in hunted bighorn sheep. *Proc. Natl Acad. Sci. USA* **111**, 13223–13228 (2014).
21. Blanchard, P., Festa-Bianchet, M., Gaillard, J.-M. & Jorgenson, J. T. Maternal condition and offspring sex ratio in polygynous ungulates: a case study of bighorn sheep. *Behav. Ecol.* **16**, 274–279 (2005).
22. Martin, J. G. A. & Festa-Bianchet, M. Sex ratio bias and reproductive strategies: what sex to produce when? *Ecology* **92**, 441–449 (2011).
23. Festa-Bianchet, M. The social system of bighorn sheep: grouping patterns, kinship and female dominance rank. *Anim. Behav.* **42**, 71–82 (1991).
24. Hewison, A. J. M. *et al.* Big mothers invest more in daughters – reversed sex allocation in a weakly polygynous mammal. *Ecol. Lett.* **8**, 430–437 (2005).
25. Hewison, A. J. M. & Gaillard, J.-M. Birth-sex ratios and local resource competition in roe deer, *Capreolus capreolus*. *Behav. Ecol.* **7**, 461–464 (1996).
26. Jones, P. H., Van Zant, J. L. & Dobson, F. S. Variation in reproductive success of male and female Columbian ground squirrels (*Urocitellus columbianus*). *Can. J. Zool.* **90**, 736–743 (2012).
27. Murie, J. O. Mating behavior of Columbian ground squirrels. I. Multiple mating by females and multiple paternity. *Can. J. Zool.* **73**, 1819–1826 (1995).
28. Neuhaus, P. & Pelletier, N. Mortality in relation to season, age, sex, and reproduction in Columbian ground squirrels (*Spermophilus columbianus*). *Can. J. Zool.* **79**, 465–470 (2001).
29. Coltman, D. W., Festa-Bianchet, M., Jorgenson, J. T. & Strobeck, C. Age-dependent sexual selection in bighorn rams. *Proc. Biol. Sci.* **269**, 165–172 (2002).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank Y. Vindenes, S. Cubaynes, S. West, R. K. Kanda, J. A. Deere, J. Barthold, M. Brouard, R. A. Pozo, and E. G. Simmonds for comments. We thank M. Festa-Bianchet and F. Pelletier for access to Bighorn sheep data and feedback. We acknowledge the use of the University of Oxford Advanced Research Computing facility. S.S. was funded by an ERC Advanced Grant to T.C., P.N. is funded by a Swiss National Science Foundation grant (SNF 3100AO-109816), and L.T. was funded by grants from the European Commission (Marie Curie Fellowship 254442) and the Carnegie Corporation of New York (B8749.R01).

Author Contributions S.S., J.M.G. and T.C. conceived and designed the study. S.S. developed the models and, with S.T., derived the formulas. S.S. and T.C. wrote the manuscript. S.S., A.G. and S.T. contributed to the mathematical formulation of the model. P.N. collated data on Columbian ground squirrels. L.T. parameterized data for bighorn sheep. T.C. parameterized data for squirrels. All authors edited the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.S. (Susanne.Schindler@zoo.ox.ac.uk).

METHODS

Data reporting. No statistical methods were used to predetermine sample size.

Summary. We use deterministic two-sex integral projection models (IPMs), that is, stochastic effects are excluded, because demographic rates of males and females typically fluctuate together with the environment. To calculate the RV for two sexes in an age- and size-structured population model we construct a sex-, age-, and size-structured generation/cohort projection matrix. The left eigenvector of this generation matrix gives the RV at birth for females and males, $v^{\circ}(s_0)$ and $v^{\sigma}(s_0)$, as a function of birth weight s_0 . Maternal body weight is a cue for maternal condition because it strongly affects survival^{30,31} and breeding success^{32,33}. We therefore weight the RV with the offspring distribution conditioned on maternal weight and age to obtain the RV of daughters (d) and sons (s) to a mother of weight s_f and age a , $v_a^d(s_f)$ and $v_a^s(s_f)$. The difference between the RV of male and female offspring is $\Delta v_a(s_f) = v_a^s(s_f) - v_a^d(s_f)$; it is a function of maternal weight and age, and if $\Delta v_a(s_f)$ is nearly linear then the slope determines whether there is a Trivers–Willard effect (type 1, positive slope) or a reversed Trivers–Willard effect (type 2, negative slope). In some instances, the association is nonlinear, leading us to identify type-3 and -4 Trivers–Willard effects (see main text).

The Model. The integral projection model uses four functions: first, the age-specific survival (superscript s) probability $p_a^s(s)$ as a function of weight s ; second, the age-specific probability distribution $p_a^s(s_2|s_1)$ of weight at next time-step s_2 as a function of current weight s_1 conditioned on survival (superscript g, growth); third, the offspring number $R_a(s)$ as a function of maternal age a and weight s ; and fourth, the offspring weight distribution $f(s|s_f, s_m)$ as a function of maternal, s_f , and paternal weight s_m at conception. Mating behaviour is described by the set of mating probabilities between every parental weight and age combination $m(a_f, s_f; a_m, s_m)$. We use models applied to Columbian ground squirrels¹⁸ and big-horn sheep²⁰. For both species, we assume that male mating probability increases with male body weight; we assume, in contrast, that female mating probability is independent of body weight once it is above a size (squirrels) or age (sheep) threshold. For sheep, we assume that male mating probability is zero for rams below 80 kg and then increases linearly with the weight of the ram. The population level offspring sex ratio at birth is unity in both models (although the observed sex ratio in the squirrel population is extremely flexible). The IPM projects a population structured along a continuous trait (here, body weight) from one time-step to the next. There are no stochastic or density effects, and the numerical iteration of the projection converges quickly. Model outputs are the population growth rate λ and asymptotic age- and weight-distribution of the female and male components of the population, $n^{\circ}(a, s)$ and $n^{\sigma}(a, s)$.

Construction of a cohort projection matrix. Age-structured models readily generalize to age- and stage-structured models³⁴. We use the asymptotic population distributions, $n^{\circ}(a, s)$ and $n^{\sigma}(a, s)$, to calculate the age- and size-specific fertility functions ($M_a^{\circ}(s)$ and $M_a^{\sigma}(s)$), the age- and size-specific inheritance functions ($D_a^{\circ}(s_0|s)$ and $D_a^{\sigma}(s_0|s)$), and the age- and birth-size-specific survivorship functions ($L_a^{\circ}(s, s_0)$ and $L_a^{\sigma}(s, s_0)$). The exact calculation of $M_a^{\circ}(s)$, $M_a^{\sigma}(s)$, $D_a^{\circ}(s_0|s)$, $D_a^{\sigma}(s_0|s)$, $L_a^{\circ}(s, s_0)$, and $L_a^{\sigma}(s, s_0)$ is given in detail below. Here we provide definitions of these terms and outline how we use them to construct a generation projection matrix.

The total number of offspring (both sexes) produced by a female (or male) of age a and weight s is $M_a^{\circ}(s)$ (and $M_a^{\sigma}(s)$ for males). The fraction of these offspring that start life with birth weight s_0 is given by $D_a^{\circ}(s_0|s)$ (and $D_a^{\sigma}(s_0|s)$ for males). Among the $M_a^{\circ}(s)$ offspring that a mother produces are $M_a^{\circ\circ}(s)$ daughters and $M_a^{\circ\sigma}(s)$ sons. Similarly, the daughters ($M_a^{\sigma\circ}(s)$) and sons ($M_a^{\sigma\sigma}(s)$) of a father sum to $M_a^{\sigma}(s)$. We assume that mothers and fathers produce sons and daughters at an even ratio, therefore $M_a^{\circ\circ}(s) = M_a^{\circ\sigma}(s) = \frac{1}{2}M_a^{\circ}(s)$ and $M_a^{\sigma\circ}(s) = M_a^{\sigma\sigma}(s) = \frac{1}{2}M_a^{\sigma}(s)$. The probability that a newborn male (or female) of age 1 and birth weight s_0 is alive with weight s at age a is $L_a^{\sigma}(s, s_0)$ (and $L_a^{\circ}(s, s_0)$ for females).

We assume that the inheritance of body weight is independent of offspring sex, that is, $D_a^{\circ}(s_0|s) = D_a^{\sigma\circ}(s_0|s) = D_a^{\circ\sigma}(s_0|s)$ and $D_a^{\sigma}(s_0|s) = D_a^{\sigma\sigma}(s_0|s) = D_a^{\sigma\circ}(s_0|s)$. With $r = \ln \lambda$, the stable population growth rate, we define the operators:

$$A_r^{\circ\circ}(s_1, s_0) = \sum_a \int e^{-ra} \frac{1}{2} D_a^{\circ}(s_1|s) M_a^{\circ\circ}(s) L_a^{\circ}(s, s_0) ds \quad (1)$$

$$A_r^{\circ\sigma}(s_1, s_0) = \sum_a \int e^{-ra} \frac{1}{2} D_a^{\circ}(s_1|s) M_a^{\circ\sigma}(s) L_a^{\circ}(s, s_0) ds \quad (2)$$

$$A_r^{\sigma\circ}(s_1, s_0) = \sum_a \int e^{-ra} \frac{1}{2} D_a^{\sigma}(s_1|s) M_a^{\sigma\circ}(s) L_a^{\sigma}(s, s_0) ds \quad (3)$$

$$A_r^{\sigma\sigma}(s_1, s_0) = \sum_a \int e^{-ra} \frac{1}{2} D_a^{\sigma}(s_1|s) M_a^{\sigma\sigma}(s) L_a^{\sigma}(s, s_0) ds \quad (4)$$

where $A_r^{\circ\circ}(s_1, s_0)$, for example, gives the fraction of male progenies with birth weight s_1 that are produced by a female of birth weight s_0 during the course of her life. We construct the generation/cohort projection matrix A :

$$A = \begin{pmatrix} A_r^{\circ\circ} & A_r^{\circ\sigma} \\ A_r^{\sigma\circ} & A_r^{\sigma\sigma} \end{pmatrix} \quad (5)$$

which projects one cohort to the next. Let $u^{\circ}(s_0)$ be the female newborn distribution and $u^{\sigma}(s_0)$ the male newborn distribution. Matrix A maps the cohort distribution of newborns of generation t to the offspring distribution produced by this cohort in generation $t + 1$. That is for a population with growth rate r :

$$\begin{pmatrix} u_{t+1}^{\circ} \\ u_{t+1}^{\sigma} \end{pmatrix} = A \begin{pmatrix} u_t^{\circ} \\ u_t^{\sigma} \end{pmatrix} \quad (6)$$

Matrix A has a dominant eigenvalue of unity, a right eigenvector that gives the stable newborn distribution, that is, $(u^{\circ}, u^{\sigma})^T$, and a left eigenvector that gives the RV, that is, (v°, v^{σ}) , of newborns of each size and sex.

Calculation of RV as a function of maternal weight. We are interested in optimal sex allocation, an optimization task conducted by mothers. Males can influence primary sex ratio, for instance, by transmitting an unequal share of male or female gametes³⁵, but the female is in principle able to render her partner's preference of offspring sex ineffective. She can do so by cryptic gamete choice or, if females are the heterogametic sex, by producing more gametes of one sex than the other.

We use the left eigenvector of matrix A which gives the female and male RV for each birth weight, $v^{\circ}(s_0)$ and $v^{\sigma}(s_0)$, and weight it with the offspring distribution of a mother of age a and weight s , $D_a^{\circ}(s_0|s)$:

$$v_a^s(s_f) = \int v^{\sigma}(s_0) D_a^{\sigma}(s_0|s_f) ds_0 \quad \text{and} \quad v_a^d(s_f) = \int v^{\circ}(s_0) D_a^{\circ}(s_0|s_f) ds_0 \quad (7)$$

to obtain the RV of a son and a daughter as a function of maternal weight.

Trivers–Willard effect, reversed Trivers–Willard effect, and other sex allocation tactics. We study the difference between male and female RV as a function of maternal weight: $\Delta v_a(s_f) = v_a^s(s_f) - v_a^d(s_f)$, where s_f is maternal weight. If $\Delta v_a(s)$ is approximately linear then the sign of the approximate slope of $\Delta v_a(s)$ determines whether the species is predicted to show a Trivers–Willard effect or a reversed Trivers–Willard effect. Positive slope, that is, female RV exceeds male RV at low maternal body weight and male RV exceeds female RV at high maternal body weight, means there is a Trivers–Willard effect (Fig. 1a). A negative slope of $\Delta v_a(s)$ implies a reversed Trivers–Willard effect (Fig. 1d):

$$\Delta v_a(s_{\max}) - \Delta v_a(s_{\min}) \begin{cases} > 0 & \text{Trivers – Willard effect (type 1),} \\ < 0 & \text{reversed Trivers – Willard effect (type 2)} \end{cases} \quad (8)$$

where s_{\max} and s_{\min} denote the midpoint of the largest and smallest reproductive weight-class, respectively.

If $\Delta v_a(s)$ is strongly nonlinear such that a linear approximation would be inappropriate, then the shape of $\Delta v_a(s)$ defines the optimal sex allocation tactic. For example, a u-shape implies that intermediate-sized mothers should produce daughters, while mothers at the extreme ends of the weight scale should produce sons; we name this tactic a type 3 Trivers–Willard effect (Fig. 1e). An n-shape implies that intermediate mothers should produce sons, while mothers at the extreme ends of the weight scale should produce daughters; we name this tactic a type 4 Trivers–Willard effect (Fig. 1f).

Sensitivity analysis. We approximate the sensitivity of Δv_a to an upward perturbation ε of any parameter p by:

$$\frac{\partial \Delta v_a}{\partial p}(s) \approx \frac{\Delta v_a(s, p(1 + \varepsilon)) - \Delta v_a(s, p)}{p\varepsilon} \quad (9)$$

We perturb survival parameters by 1% downwards and all other parameters by 1% upwards, which means that any perturbation results in increased mortality, growth, or inheritance probabilities. If $\Delta v_a(s)$ and $\frac{\partial \Delta v_a(s)}{\partial p}$ are reasonably linear

then a change in the slope implies either a strengthening or weakening effect. If $\frac{\partial \Delta v_a(s)}{\partial p}$ has a steeper slope of the same positive (or negative) sign as $\Delta v_a(s)$ then we say that a parameter perturbation strengthens a Trivers–Willard effect (Fig. 1b; or reversed Trivers–Willard effect). If the slope of $\frac{\partial \Delta v_a(s)}{\partial p}$ is shallower than that of $\Delta v_a(s)$, then we speak of weakening the effect (Fig. 1c). If the slope of $\frac{\partial \Delta v_a(s)}{\partial p}$ is negative in contrast to a positive slope of $\Delta v_a(s)$, then a Trivers–Willard effect has been reversed (Fig. 1d). However, neither $\Delta v_a(s)$ nor $\frac{\partial \Delta v_a(s)}{\partial p}$ have to be linear.

For example, when we perturb the male survival intercept in the squirrel model, a reversed Trivers–Willard effect changes into a type 3 effect (u-shape). The results of the sensitivity analyses for squirrels in are shown in Extended Data Figs 4 and 5, and for sheep in Extended Data Figs 7 and 8.

Mean and variance in RV. To calculate the properties of the distribution of offspring's RV we use the stable population distribution $n^{\circ}(a, s)$ and $n^{\sigma}(a, s)$ from iterating the two-sex IPM. We denote the mean and variance of the RV distribution with $\mathbb{E}(v)$ and $\text{Var}(v)$ and calculate it by

$$\mathbb{E}(v_a^d) = \frac{\int_0^{\infty} v_a^d(s) n^{\circ}(a, s) ds}{\int_0^{\infty} n^{\circ}(a, s) ds} \quad \text{and} \quad \mathbb{E}(v_a^s) = \frac{\int_0^{\infty} v_a^s(s) n^{\sigma}(a, s) ds}{\int_0^{\infty} n^{\sigma}(a, s) ds} \quad (10)$$

$$\text{Var}(v_a^d) = \mathbb{E}((v_a^d)^2) - (\mathbb{E}(v_a^d))^2 \quad \text{and} \quad \text{Var}(v_a^s) = \mathbb{E}((v_a^s)^2) - (\mathbb{E}(v_a^s))^2 \quad (11)$$

Calculation of D_a . The term $D_a^{\sigma}(s_0|s)$ (or $D_a^{\circ}(s_0|s)$) denotes the probability that an offspring produced by a father (or a mother, respectively) of age a and weight s is born with birth weight s_0 . The terms $D_a^{\sigma}(s_0|s)$ and $D_a^{\circ}(s_0|s)$ are conditional probabilities, that is, $\int D_a^{\sigma}(s_0|s) ds_0 = \int D_a^{\circ}(s_0|s) ds_0 = 1$. They are calculated using the mating function $m(a_f, s_f; a_m, s_m)$ and fertility function $R_{a_f}(s_f)$ of the two-sex IPM, as well as the stable female and male distributions $n^{\circ}(a_f, s_f)$ and $n^{\sigma}(a_m, s_m)$, which are obtained by iteration. The D_a functions are calculated by:

$$D_{a_f}^{\circ}(s_0|s_f) = \begin{cases} \frac{\sum_{a_m} \int f(s_0|s_f, s_m) m(a_f, s_f; a_m, s_m) R_{a_f}(s_f) n^{\sigma}(a_m, s_m) ds_m}{\sum_{a_m} \int m(a_f, s_f; a_m, s_m) R_{a_f}(s_f) n^{\sigma}(a_m, s_m) ds_m} & \text{if } \sum_{a_m} \int m(a_f, s_f; a_m, s_m) R_{a_f}(s_f) n^{\sigma}(a_m, s_m) ds_m > 0 \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

$$D_{a_m}^{\sigma}(s_0|s_m) = \begin{cases} \frac{\sum_{a_f} \int f(s_0|s_f, s_m) m(a_f, s_f; a_m, s_m) R_{a_f}(s_f) n^{\circ}(a_f, s_f) ds_f}{\sum_{a_f} \int m(a_f, s_f; a_m, s_m) R_{a_f}(s_f) n^{\circ}(a_f, s_f) ds_f} & \text{if } \sum_{a_f} \int m(a_f, s_f; a_m, s_m) R_{a_f}(s_f) n^{\circ}(a_f, s_f) ds_f > 0 \\ 0 & \text{otherwise} \end{cases} \quad (13)$$

Calculation of M_a . The term $M_a^{\circ\circ}(s)$ denotes the total number of daughters produced by a mother of weight s and age a . Similarly, $M_a^{\sigma\sigma}(s)$ gives the number of sons to a mother, $M_a^{\circ\sigma}(s)$ gives the number of daughters to a father, and $M_a^{\sigma\sigma}(s)$ gives the number of sons to a father of age a and weight s . The terms are calculated using the mating function $m(a_f, s_f; a_m, s_m)$ and fertility function $R_{a_f}(s_f)$ of the two-sex IPM, as well as the stable female and male distributions $n^{\circ}(a_f, s_f)$

and $n^{\sigma}(a_m, s_m)$, which are obtained by iteration. The M_a° and M_a^{σ} functions are calculated by:

$$M_{a_f}^{\circ}(s_f) = C \sum_{a_m} \int m(a_f, s_f; a_m, s_m) R_{a_f}(s_f) n^{\sigma}(a_m, s_m) ds_m \quad (14)$$

$$M_{a_m}^{\sigma}(s_m) = C \sum_{a_f} \int m(a_f, s_f; a_m, s_m) R_{a_f}(s_f) n^{\circ}(a_f, s_f) ds_f \quad (15)$$

with the normalization constant

$$C = \frac{\sum_{a_f > a_{\min}} \int_{s_{\min}} n^{\circ}(a_f, s_f) ds_f}{\sum_{a_f, a_m} \int m(a_f, s_f; a_m, s_m) n^{\circ}(a_f, s_f) n^{\sigma}(a_m, s_m) ds_f ds_m} \quad (16)$$

where a_{\min} and s_{\min} are the minimum age and weight necessary to start reproducing.

Calculation of L_a . The term $L_a^{\circ}(s, s_0)$ (or $L_a^{\sigma}(s, s_0)$) gives the probability that a newborn female (or a newborn male, respectively) born with weight s_0 will be alive at age a and weigh s weight units. The L_a terms are calculated recursively using the combined and age-specific survival and growth function p_a^{sg} (see refs 18 and 20 for more detail on p_a^{sg}) by

$$L_1^{\circ}(s, s_0) = L_1^{\sigma}(s, s_0) = \begin{cases} 1 & \text{if } s = s_0 \\ 0 & \text{otherwise} \end{cases} \quad (17)$$

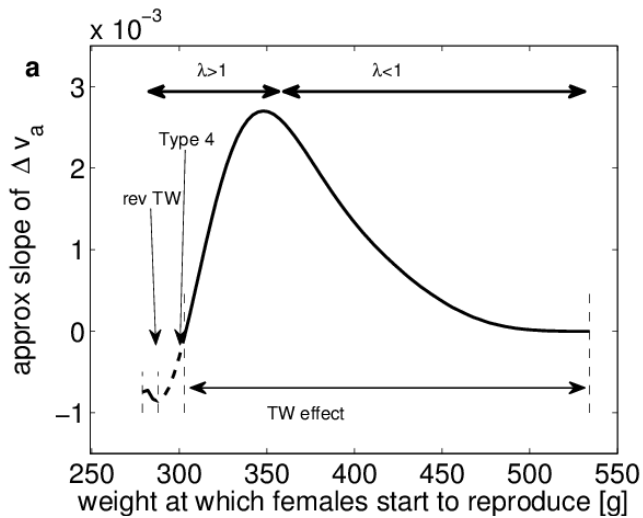
$$L_{a_f}^{\circ}(s, s_0) = \int p_{a_f-1}^{\text{sg}, \circ}(s, s_1) L_{a_f-1}^{\circ}(s_1, s_0) ds_1 \quad \text{for } a_f \geq 2 \quad (18)$$

$$L_{a_m}^{\sigma}(s, s_0) = \int p_{a_m-1}^{\text{sg}, \sigma}(s, s_1) L_{a_m-1}^{\sigma}(s_1, s_0) ds_1 \quad \text{for } a_m \geq 2 \quad (19)$$

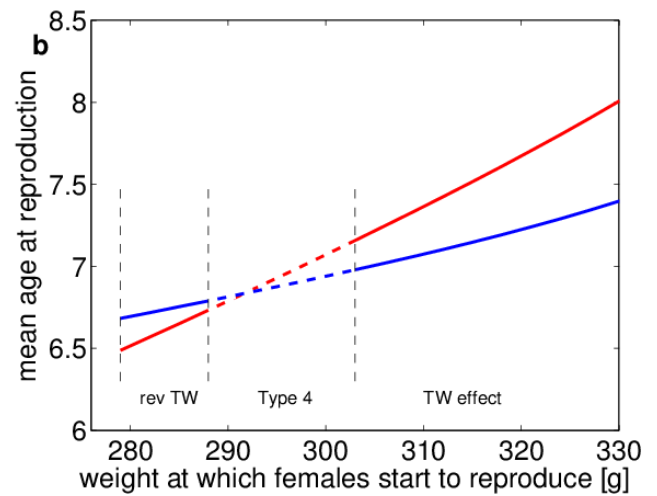
where $p_a^{\text{sg}, \circ}(s, s_1)$ and $p_a^{\text{sg}, \sigma}(s, s_1)$ give the probability of a female or male, respectively, of age a and weight s_1 to survive and attain weight s in the next time step.

Code availability. A Matlab script for calculating RV as left eigenvectors from the cohort projection matrix is given in Supplementary Information, appendix B. The script covers calculations in equations (1–5, 7, and 12–19) and does not include the two IPM models described elsewhere^{18,20}.

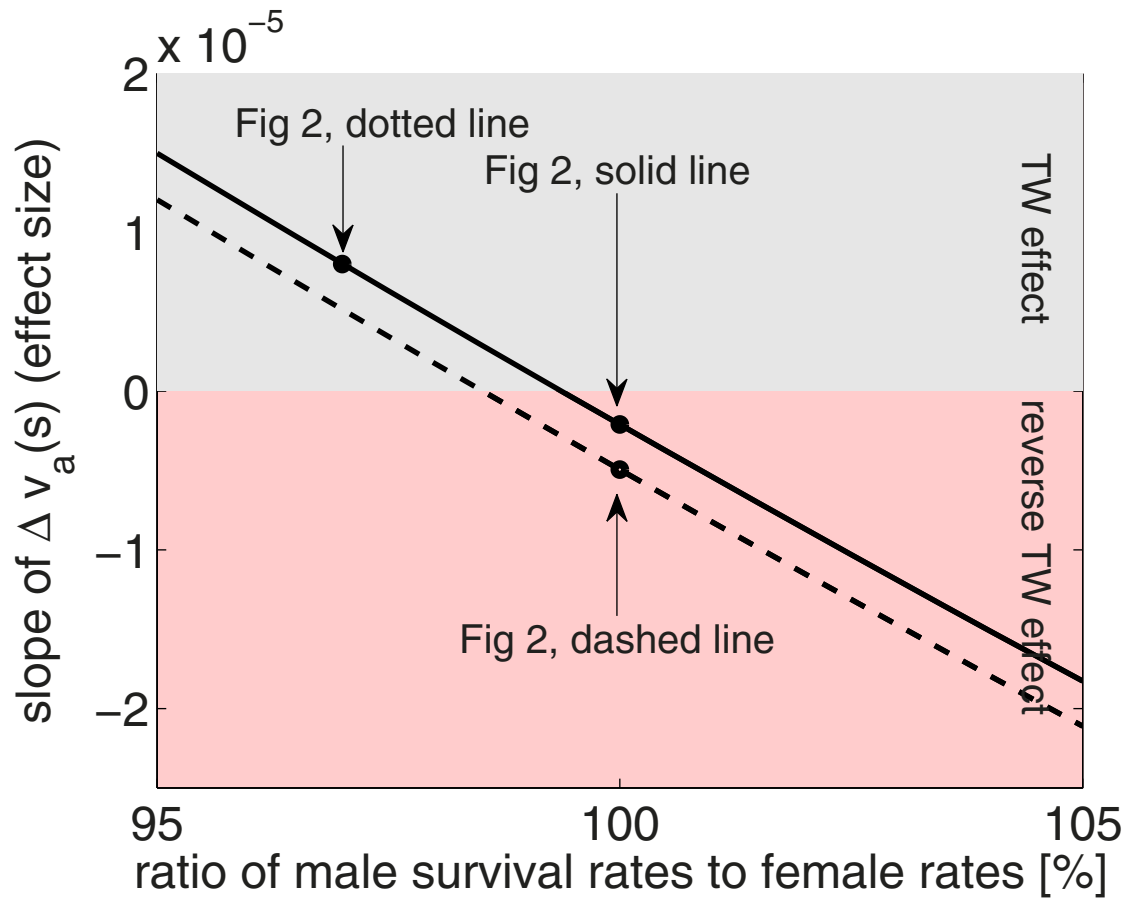
- Neuhaus, P. Weight comparisons and litter size manipulation in Columbian ground squirrels (*Spermophilus columbianus*) show evidence of costs of reproduction. *Behav. Ecol. Sociobiol.* **48**, 75–83 (2000).
- Nussey, D. H. *et al.* Patterns of body mass senescence and selective disappearance differ among three species of free-living ungulates. *Ecology* **92**, 1936–1947 (2011).
- Bronson, M. T. Altitudinal variation in the life history of the golden-mantled ground squirrel (*Spermophilus lateralis*). *Ecology* **60**, 272–279 (1979).
- Festa-Bianchet, M., Gaillard, J. & Jorgenson, J. Mass- and density-dependent reproductive success and reproductive costs in a capital breeder. *Am. Nat.* **152**, 367–379 (1998).
- Steiner, U. K., Tuljapurkar, S. & Coulson, T. Generation time, net reproductive rate, and growth in stage-age-structured populations. *Am. Nat.* **183**, 771–783 (2014).
- Edwards, A. M. & Cameron, E. Z. Forgotten fathers: paternal influences on mammalian sex allocation. *Trends Ecol. Evol.* **29**, 158–164 (2014).
- McGraw, J. B. & Caswell, H. Estimation of individual fitness from life-history data. *Am. Nat.* **147**, 47–64 (1996).



Extended Data Figure 1 | Mean female age at reproduction affects optimal sex allocation. **a**, Slope of the difference between male and female RV as a function of the size threshold above which females reproduce. The male size threshold is fixed at 279 g. Negative values indicate a reversed Trivers–Willard effect, positive values a Trivers–Willard effect. Dashed lines indicate a type 4 effect. When the population growth rate λ is greater than 1 (growing population), increasing female age at reproduction selects towards a Trivers–

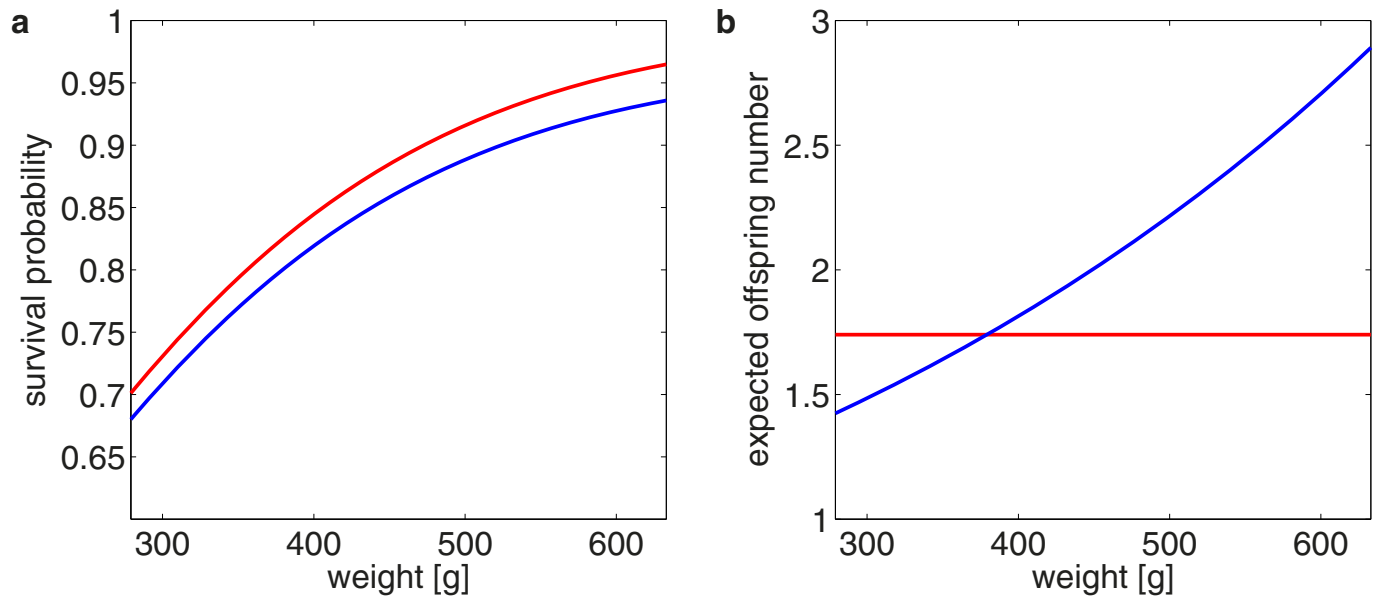


Willard effect. In contrast, when the population is shrinking ($\lambda < 1$), reproducing at a later age increases fitness³⁶ and selects towards a reversed Trivers–Willard effect with increasing female age at reproduction. **b**, Mean maternal (red) and paternal (blue) age at reproduction as a function of the size threshold at which females reproduce. Dashed lines indicate the range of size thresholds that cause a type 4 effect.

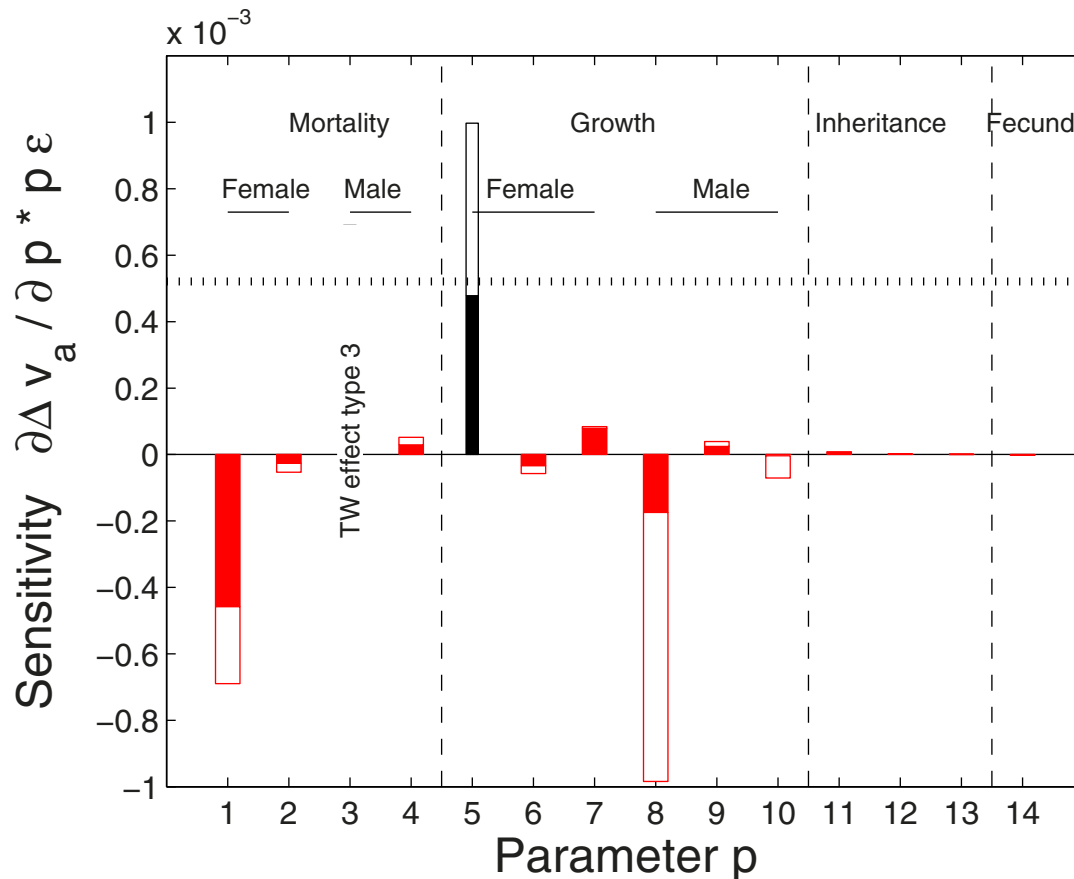


Extended Data Figure 2 | Strength of Trivers–Willard and reversed Trivers–Willard effects in squirrels as a function of the male to female survival ratio. The x -axis plots the ratio of male to female survival rate (independent of size and age) to the slope of the difference between male and female reproductive value, $\Delta v_a(s)$. Positive values indicate a Trivers–Willard effect (grey background), negative values a reversed Trivers–Willard effect (red

background). The more positive (or negative) the slope of $\Delta v_a(s)$ the more the expected sex ratio in offspring to good-condition mothers is biased towards males (or females). Solid line, no sex differences in mortality; dashed line, strength of mate selection has been increased from $\rho = 0.1$ to $\rho = 0.25$ (see Supplementary Table 1). Points highlighted with arrows indicate the settings that are used in Fig. 2 to plot $\Delta v_a(s)$ against maternal size s .

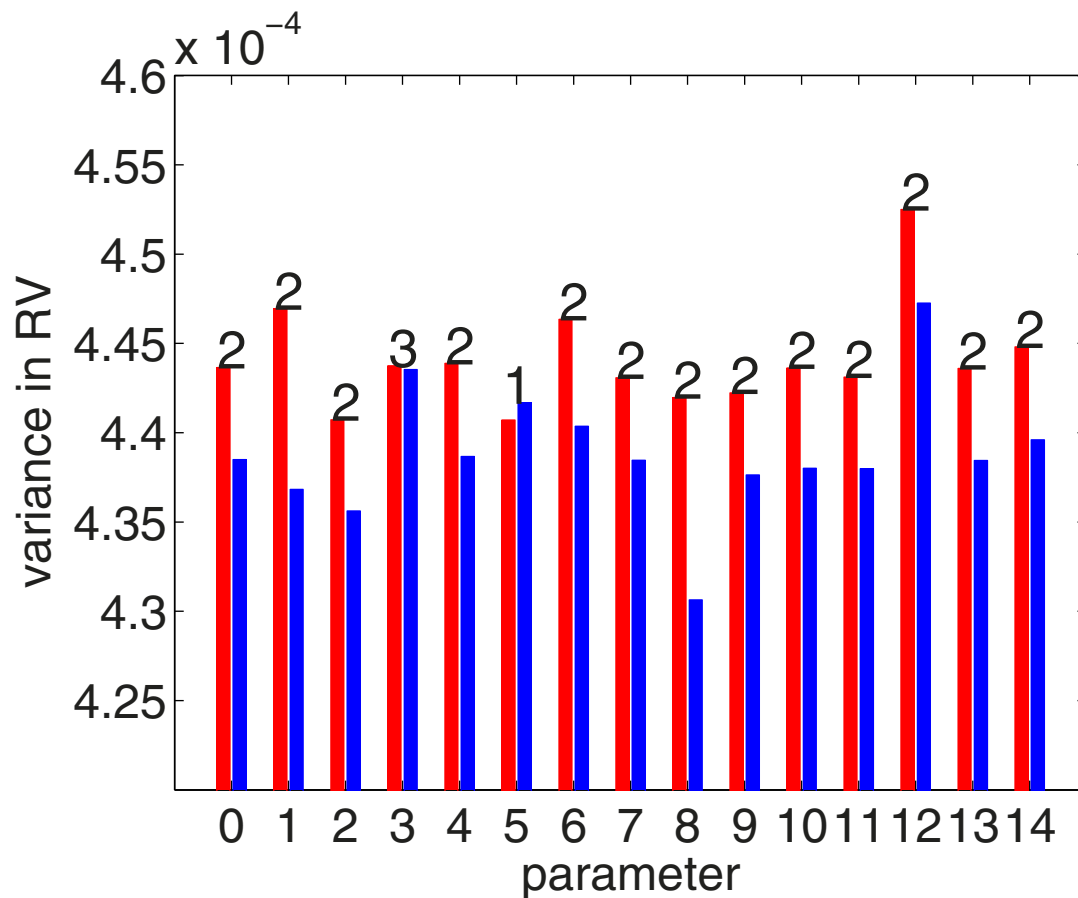


Extended Data Figure 3 | Trade-off between survival and reproduction in squirrels. **a**, Females (red) have higher survival rates than males (blue) at all ages. **b**, Small females are expected to produce more offspring than small males, while large females produce less offspring than large males.



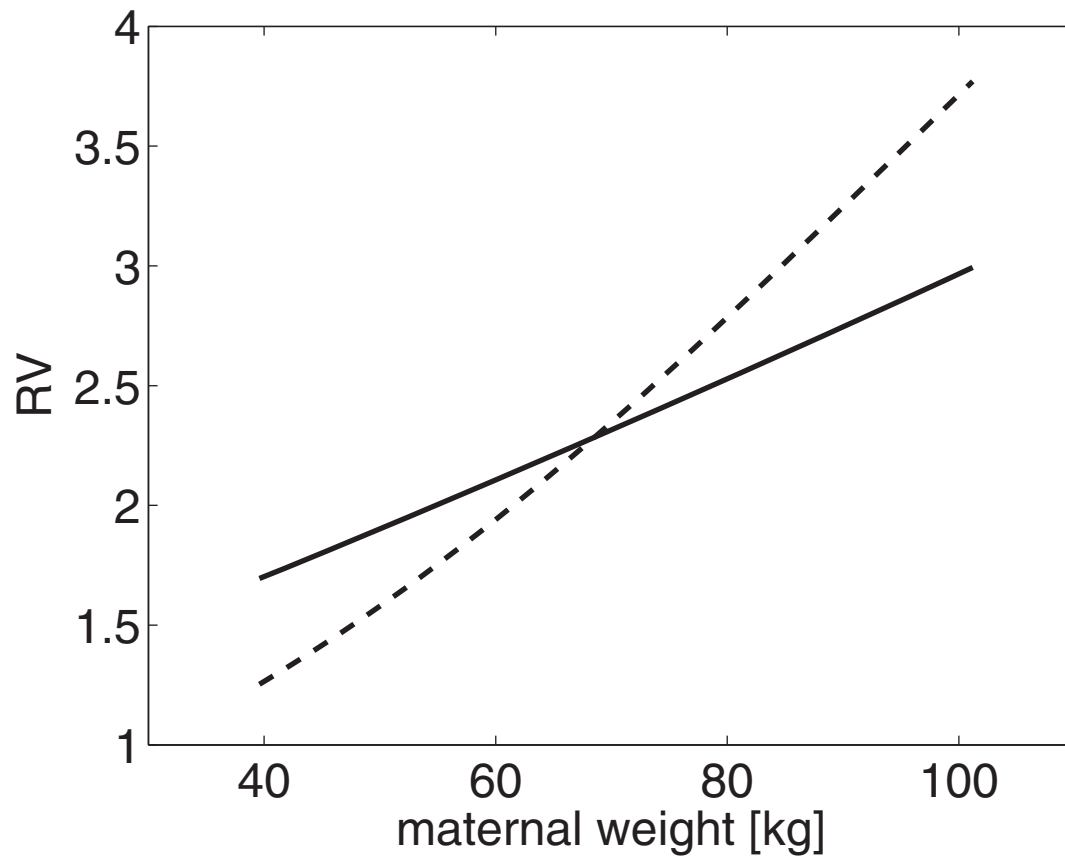
Extended Data Figure 4 | Sensitivity of the reversed Trivers-Willard effect to parameter perturbations in squirrels. When bars lie in the positive (or negative) range, then a change in parameters works towards a Trivers-Willard effect (or strengthening the reversed Trivers-Willard effect). The horizontal dashed line shows the difference in slope needed to neutralise the reversed Trivers-Willard effect. The bar above the dashed line indicates a Trivers-Willard effect (black); bars below indicate a reversed Trivers-Willard effect (red). Filled fractions of the bars indicate the contribution of change caused by parameter perturbation owing to change of female RV, and, in the white fractions, to change in male RV (see also Extended Data Fig. 5). Bars 1 to 4 show the sensitivity of the Trivers-Willard effect in squirrels to perturbations of the following parameters by 1% downwards (which corresponds to higher

mortality in the sex affected): (1) female survival intercept; (2) female survival slope; (3) male survival intercept, parameter change resulted in curved Δv_a which we indicate with 'TW effect type 3' and omit the bar; (4) male survival slope. Bars from 5 to 14 show the sensitivity of the Trivers-Willard effect in squirrels to perturbations of the following parameters by 1% upwards (which corresponds to higher rates in the affected sex): (5) female growth (mean intercept); (6) female growth (mean slope); (7) female growth variance; (8) male growth (mean intercept); (9) male growth (mean slope); (10) male growth variance; (11) inheritance (mean intercept); (12) inheritance (mean slope); (13) inheritance variance; (14) expected offspring number. All parameters are listed in Supplementary Table 1.



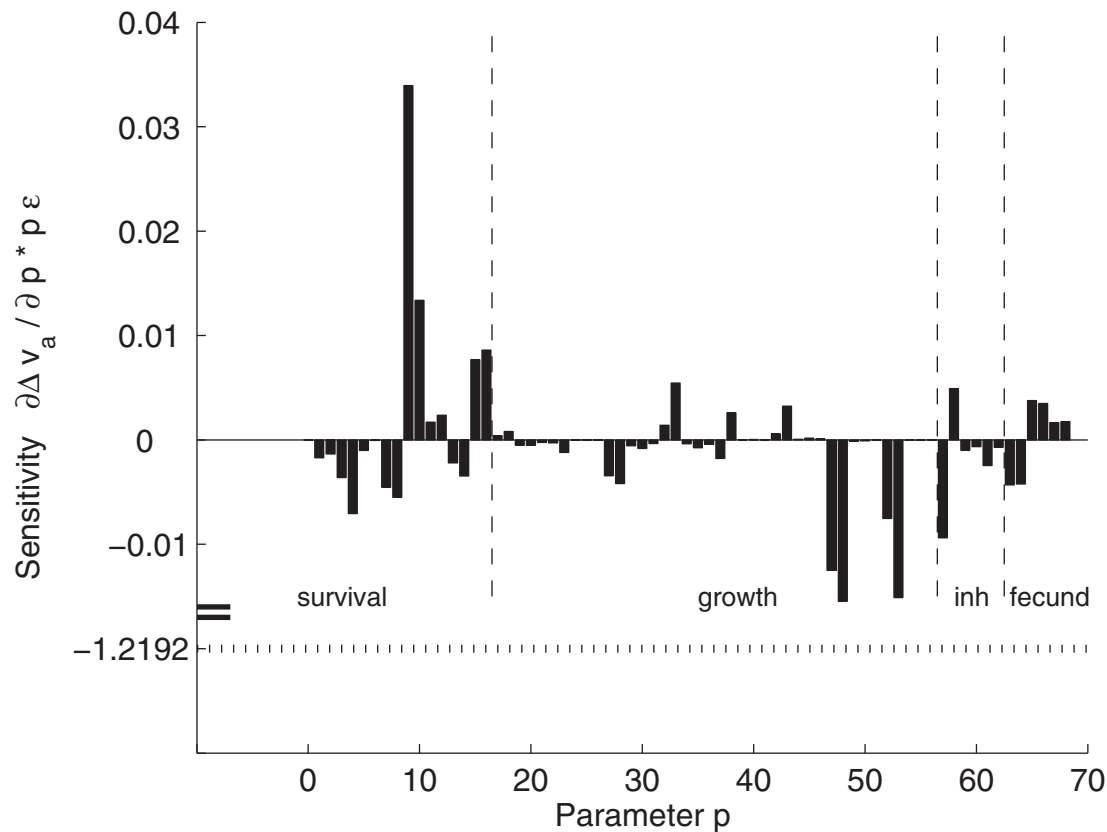
Extended Data Figure 5 | Female (red) and male (blue) variance in RV in original model (bar 0) and when parameters are perturbed (bars 1–14) in squirrels. Number 1 above bars indicates a Trivers–Willard effect, number 2 a reversed Trivers–Willard effect, and number 3 a type 3 Trivers–Willard effect. Bars 1 to 4 show variances in RV when the survival parameters are perturbed by 1% downwards (which corresponds to higher mortality in the affected sex): (1) female survival intercept; (2) female survival slope; (3) male survival intercept; and (4) male survival slope. Bars 5 to 14 show variances in RV when

the following parameters are perturbed by 1% upwards (which corresponds to higher rates in the affected sex): (5) female growth (mean intercept); (6) female growth (mean slope); (7) female growth variance; (8) male growth (mean intercept); (9) male growth (mean slope); (10) male growth variance; (11) inheritance (mean intercept); (12) inheritance (mean slope); (13) inheritance variance; and (14) expected offspring number. All parameters are listed in Supplementary Table 1.



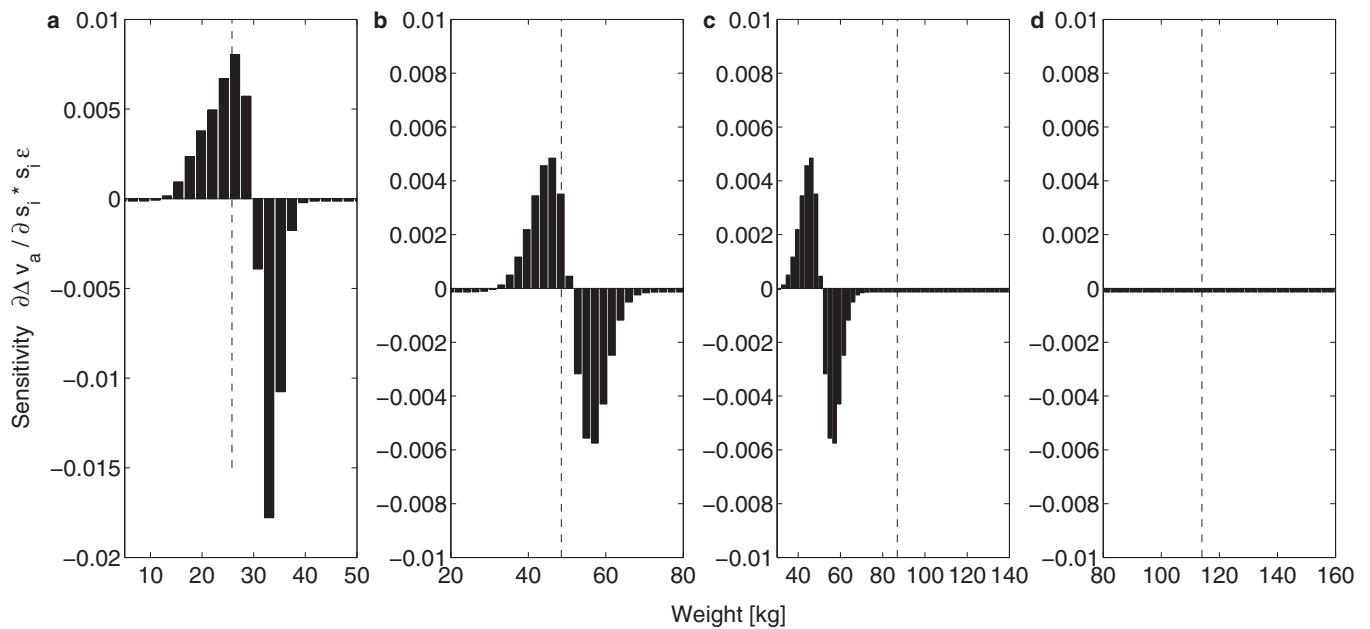
Extended Data Figure 6 | Reproductive value (RV) of female (solid line) and male (dashed line) offspring in bighorn sheep. For small mothers, daughters have higher RV than sons. For large mothers, sons have higher RV

than daughters. We scaled the RV of females and males such that the female RV of the smallest reproductive size class is 1.



Extended Data Figure 7 | Sensitivity of the Trivers-Willard effect to parameter perturbations in sheep. Bars in the positive range indicate that the Trivers-Willard effect is strengthened; bars in the negative range indicate a weakened Trivers-Willard effect. The horizontal dotted line marks the sensitivity needed to reverse the Trivers-Willard effect. The parameters are: 1–8 female survival (2 parameters each for the stages lamb, yearling, adult,

and senescent); 9–16 male survival (2 parameters for each stage); 17–36 female growth (5 parameters for each stage); 37–56 male growth (5 parameters for each stage); 57–60 female inheritance (inh) (2 intercepts for mean and variance, 2 slopes for female contribution to mean and variance); 61–62 male inheritance (male contributions to mean and variance); and 63–68 fecundity (fecund). All parameters are listed in Supplementary Tables 2 and 3.



Extended Data Figure 8 | Sensitivity of Trivers–Willard effect to size-specific male mortality increases of 1% in sheep. a–d, The survival probability of each size class in each stage (lamb (a), yearling (b), adult (c), and senescent (d)) has been independently lowered by 1%. The vertical dashed black line denotes the mean body weight of male sheep in the corresponding stage. In the early stages (a and b, lamb and yearling) we find that male-mortality

increases in small size classes strengthen the Trivers–Willard effect, whereas male-mortality increases in heavy size classes weaken the Trivers–Willard effect. In the later stages (c and d, adult and senescent), mortality increases hardly affect the Trivers–Willard effect (note that adult rams usually weigh above 60 kg).

A novel locus of resistance to severe malaria in a region of ancient balancing selection

Malaria Genomic Epidemiology Network*

The high prevalence of sickle haemoglobin in Africa shows that malaria has been a major force for human evolutionary selection, but surprisingly few other polymorphisms have been proven to confer resistance to malaria in large epidemiological studies^{1–3}. To address this problem, we conducted a multi-centre genome-wide association study (GWAS) of life-threatening *Plasmodium falciparum* infection (severe malaria) in over 11,000 African children, with replication data in a further 14,000 individuals. Here we report a novel malaria resistance locus close to a cluster of genes encoding glycoporphins that are receptors for erythrocyte invasion by *P. falciparum*. We identify a haplotype at this locus that provides 33% protection against severe malaria (odds ratio = 0.67, 95% confidence interval = 0.60–0.76, P value = 9.5×10^{-11}) and is linked to polymorphisms that have previously been shown to have features of ancient balancing selection, on the basis of haplotype sharing between humans and chimpanzees⁴. Taken together with previous observations on the malaria-protective role of blood group O^{1–3,5}, these data reveal that two of the strongest GWAS signals for severe malaria lie in or close to genes encoding the glycosylated surface coat of the erythrocyte cell membrane, both within regions of the genome where it appears that evolution has maintained diversity for millions of years. These findings provide new insights into the host–parasite interactions that are critical in determining the outcome of malaria infection.

In the discovery phase of this study, we analysed GWAS data on 5,633 children with severe malaria and 5,919 population controls from The Gambia, Kenya and Malawi, and in the replication phase we analysed candidate single-nucleotide polymorphisms (SNPs) in a further 13,946 case–control samples from Burkina Faso, Cameroon, The Gambia, Ghana, Malawi, Mali and Tanzania (Extended Data Fig. 1). The majority of samples used in the discovery phase have been analysed previously by lower-resolution GWAS methods³. For this analysis we improved resolution by directly typing all samples at approximately 2.5 million SNPs using the Illumina Omni2.5M platform, followed by quality control (Extended Data Figs 1 and 2) and imputation of genotypes at over 10 million SNPs using haplotype data from the 1000 Genomes Project⁶. Imputation performance varied across populations, with accuracy varying as a function of the similarity between study and reference individuals (Extended Data Fig. 3). When testing for genetic association, principal components analysis was used to correct for population structure (Extended Data Fig. 4a–e), which reflected both geography and self-reported ethnicity. Similar results were obtained using a mixed-model approach (Extended Data Fig. 4f).

To assess the evidence for association in the discovery phase we used an approach that allows for heterogeneity in the protective effect of an allele across different study sites. This could be particularly important in our data, as high levels of genetic and ethnic diversity in Africa can result in variable patterns of linkage disequilibrium between study sites that can complicate GWAS analysis⁷. Other potential sources of heterogeneity include allelic heterogeneity and multiple independent origins of malaria resistance loci, as has been well documented at the *HBB* locus^{1,3}, as well as the high levels of genetic diversity in the

parasite⁸. Specifically, we used a Bayesian approach that combines evidence across multiple models of association by specifying a prior probability on the size and similarity of the genetic effect across populations, as well as the mode of inheritance¹. A single statistical summary of the signal of association was obtained by averaging the evidence across models, weighting each by its prior probability, and comparing the evidence to the null model of no association (model-averaged Bayes factor (BF_{avg})). Having observed the data, a posterior probability was assigned to each model, conditional on it being a true association and the model assumptions, which are described in Methods and Extended Data Fig. 5. We replicated previously reported GWAS signals^{2,3,9} at the *HBB* ($BF_{\text{avg}} = 5.8 \times 10^{24}$), *ABO* ($BF_{\text{avg}} = 6.7 \times 10^9$) and *ATP2B4* ($BF_{\text{avg}} = 4.4 \times 10^5$) loci, and a detailed analysis of key variants at these and other previously reported loci is presented elsewhere¹. A previously reported association near the gene *MARVELD3* (ref. 2) did not replicate in this data set (Supplementary Note 1). Genome-wide patterns of association with severe malaria at the 34 regions of the genome containing a variant with either a Bayes factor for the most probable model ($BF_{\text{max}} > 2.5 \times 10^4$ or with a $BF_{\text{avg}} > 2.5 \times 10^3$ are summarized in Extended Data Fig. 6 and Supplementary Table 1. Details of the evidence for association in these regions can be viewed online at <http://www.malariagen.net/resource/14>.

These data provide a rich resource of new candidate loci for further investigation. Here we focus on a region of chromosome 4 shown in Fig. 1, where the strongest signal of association (at SNP rs184895969) is located between the gene *FREM3* and a cluster of three glycoporphin genes (*GYPE*, *GYPB* and *GYPA*). Glycoporphins are sialoglycoproteins that are abundantly expressed in the erythrocyte membrane, providing a hydrophilic surface coat that is necessary for erythrocytes to flow freely in the circulation. A complex system of single-nucleotide and structural variants in this region determine the MNS blood group system¹⁰. These genes have a functional role in invasion of erythrocytes by *P. falciparum*. Glycoporphin A is the receptor for the *P. falciparum* erythrocyte-binding ligand EBA-175 (ref. 11), and glycoporphin B is a receptor for the parasite ligand EBL-1 (ref. 12). To follow up this observation, selected SNPs at this locus were genotyped by Sequenom iPLEX MassArray in the discovery and replication sample sets outlined earlier (Fig. 1 and Extended Data Fig. 7a). The combined data set of 25,498 samples provided convincing evidence of association at rs186873296 by standard fixed-effect meta-analysis ($P = 9.5 \times 10^{-11}$) as well as by the Bayesian approach described earlier ($BF_{\text{overall}} = 1.3 \times 10^8$; Fig. 2 and Methods). The derived (non-ancestral) allele of rs186873296 was at higher frequency in East Africa than in West Africa, and the greatest evidence of association was seen in Kenya, where the allele was most common with a frequency of approximately 10%. Using only replication data to avoid winner's curse, and assuming an additive model, we estimate that carrying one copy of the derived allele reduces the risk of severe malaria by about 40% in Kenya (odds ratio (OR) = 0.60, 95% confidence interval (CI) = 0.46–0.79), with a slightly smaller effect across all populations (OR = 0.67, 95% CI 0.56–0.80 in frequentist fixed-effect meta-analysis). Further details are given in Supplementary Note 2.

*Lists of participants and their affiliations appear at the end of the paper.

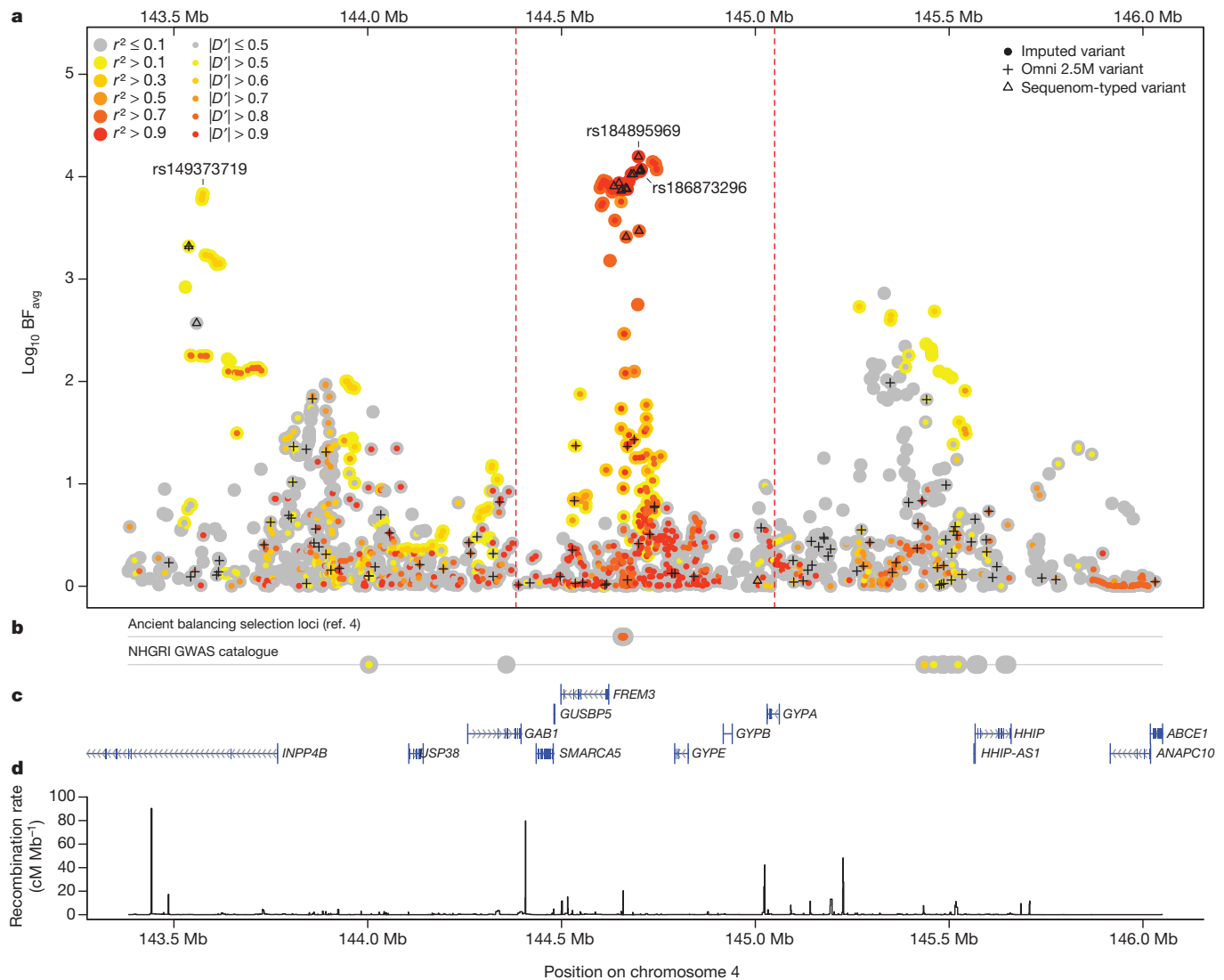


Figure 1 | Signal of association with severe malaria across the *FREM3*/*GYPE* region. **a**, Evidence for association ($\text{log}_{10} \text{BF}_{\text{avg}}$) in the discovery data. Black plus signs denote SNPs that were directly typed, and black triangles denote SNPs selected for typing on the Sequenom platform. Dotted red vertical lines indicate a region of $0.25 \text{ cM} \pm 25 \text{ kb}$ centred at the lead SNP (rs184895969). Coloured circles denote the correlation (outer circles) and $|D'|$ (inner circles)

The glycoprotein gene cluster has a complex pattern of gene conversion and structural variation that has been previously noted; indeed, it has been proposed that selective pressure due to pathogens, including malaria, has contributed to shaping diversity in this region^{10,13–16}. Using the human reference sequence and mapped sequence read data from the 1000 Genomes Project, we identified the boundaries of a 350 kb region of sequence homology surrounding these genes as well as a set of segregating gene deletions (Extended Data Fig. 8). The lead imputed marker (rs184895969) is located within 10 kb of this complex region. Imputation accuracy within the region is low using current reference data, so it is possible that the causal variant lies within the glycoprotein genes themselves but that this is obscured in the current imputed data set. With this caveat, we computed a credible set of putatively causal variants in the region; this set includes both the lead imputed marker and a linked missense mutation (rs181620317) in *FREM3* (Extended Data Fig. 7b). We note that the protective allele at rs184895969 was associated with increased *GYP A* transcription in published gene expression data for HapMap lymphoblastoid cell lines¹⁷ ($P = 0.016$; Extended Data Fig. 9); other regional analyses are described in

with rs184895969 in controls, computed from imputed haplotypes.

b, Polymorphisms shared between humans and chimpanzees, and previously reported associations with other phenotypes. NHGRI, National Human Genome Research Institute. **c**, **d**, Genes in the region and the HapMap combined recombination rate.

Supplementary Notes 1 and 2. Improved African genome variation reference panels^{7,18} are needed to understand the complex patterns of variation in this region so that the causal variant of malaria resistance can be fine-mapped with greater confidence.

A striking feature of these data is that all of the loci that reach conventional criteria for genome-wide significance ($P < 5 \times 10^{-8}$ using a fixed-effects model) are in or near genes that are important for erythrocyte function (*HBB*, *ABO*, *ATP2B4*, *FREM3*/*GYPE* region), the primary host cell of *P. falciparum*. Other erythrocyte-related genes are identified in the discovery phase analysis but do not reach genome-wide significance, including *EPB41*, which encodes erythrocyte membrane protein band 4.1 and has been implicated as a possible receptor for *P. falciparum* invasion¹⁹ (rs2985337: $\text{BF}_{\text{avg}} = 3,443$; fixed-effect meta-analysis OR = 1.16, 95% CI = 1.09–1.23, $P = 1.2 \times 10^{-6}$; see Extended Data Fig. 6).

The *ABO* locus contains a number of polymorphisms that are shared between humans and other primates, and recent analyses of sequence variation across species indicated that some of these are ancient polymorphisms that have been maintained by balancing

selection over millions of years²⁰. The current findings are of particular interest since the *FREM3/GYPE* region is one of the most prominent examples of putative ancient balancing selection in a genome-wide analysis of haplotype sharing between humans and chimpanzees⁴. The peak GWAS signal at this locus is less than 45 kb away from the shared human–chimpanzee haplotypes, which fall within the region covered by the credible set (Extended Data Fig. 7b), although they do not exhibit a strong association with severe malaria (Supplementary Note 3). To explore the genealogical relationship between the putative ancient balanced polymorphisms (ABPs) and SNPs associated with severe malaria in the *FREM3/GYPE* region, we inferred an ancestral tree²¹ from the African (YRI + LWK) part of the 1000 Genomes Project data and used it to order haplotypes in the region, labelled with the positions of ABPs, malaria-associated SNPs, and other variants of interest (Fig. 3). All three haplotypes carrying the protective allele at the directly typed marker with most evidence of association (rs186873296) carry the minor allele at the ABP markers ($D' = 1$, $P = 0.017$). By inferring the positions of putative causal mutations on the estimated genealogical tree²¹ at the lead imputed marker (rs184895969) we found evidence for a single protective mutation in Kenya ($\log_{10} \text{BF} = 3.09$; $\text{OR} = 0.6$) estimated to lie on the branch ancestral to the protective allele at the lead marker. Although the most likely position for an additional mutation was on the branch ancestral to the ABPs, a single haplotype explains most of the signal of association in this region.

These observations raise the question of whether malaria resistance loci are more likely to be found in regions of the genome that show evidence of ancient balancing selection. We therefore analysed the relationship between the regions of association in our GWAS and 125 regions of the genome found previously⁴ to contain haplotypes shared between humans and chimpanzees. The SNPs defining these haplotypes (ABPs) were not themselves enriched for association with severe malaria ($P > 0.1$, Methods). We used a simulation approach to assess the physical proximity of ABPs to the peak of association within the 34 strongest regions of association (tier 1) and 73 weaker signals (tier 2), and observed a significant relationship with tier 1 over a range of length scales (Extended Data Fig. 10a, d). We also identified the nearest gene to the lead marker within each association region as well as the gene nearest to each ABP haplotype, and performed a gene-based test for genome-wide enrichment. Strong evidence for enrichment was seen at tier 1 loci ($\text{OR} 41.0$; $P = 4 \times 10^{-7}$ by Fisher's exact test, $P_{\text{sim}} = 5 \times 10^{-4}$ using a simulation approach described in Methods, and $P = 1 \times 10^{-4}$ using the INRICH algorithm²²; Extended Data Fig. 10b–d) and a weaker trend at tier 2 loci ($\text{OR} 7.7$, $P_{\text{sim}} = 0.15$). Apart from *ABO* and *FREM3/GYPE*, there were six other GWAS loci (four in tier 1, two in tier 2) where the nearest gene to the lead marker was also the nearest gene to an ABP haplotype (*DSCAM*, *NRG1*, *CNTNAP5*, *TMEM132C*, *CACNA2D1*, *RYR2*). Although the current association evidence at these loci does not satisfy conventional criteria for genome-wide significance and they should be regarded as putative until convincingly replicated, it is noteworthy that they are all involved in key aspects of membrane biology (Supplementary Note 3).

In the largest genetic association study of malaria to date, we have identified a new locus of resistance to severe malaria that lies next to a cluster of glycoprotein genes involved in erythrocyte invasion by *P. falciparum*, and that also overlaps a locus of putative ancient balancing selection identified by analysis of haplotype sharing between humans and chimpanzees. It is possible that malaria is not the cause of the ancient balancing selection, or that it is just one of a number of opposing evolutionary driving forces, as at *ABO*, where blood group O reduces the risk of severe malaria but increases the risk of severe cholera²³. Nonetheless, these new findings raise the intriguing question of whether natural selection on malaria susceptibility has been shaping genetic diversity in humans and their ancestors for millions of years. *P. falciparum* is closely related to the chimpanzee parasite *P. reichenowi* and other parasites of African great apes^{24–26}.

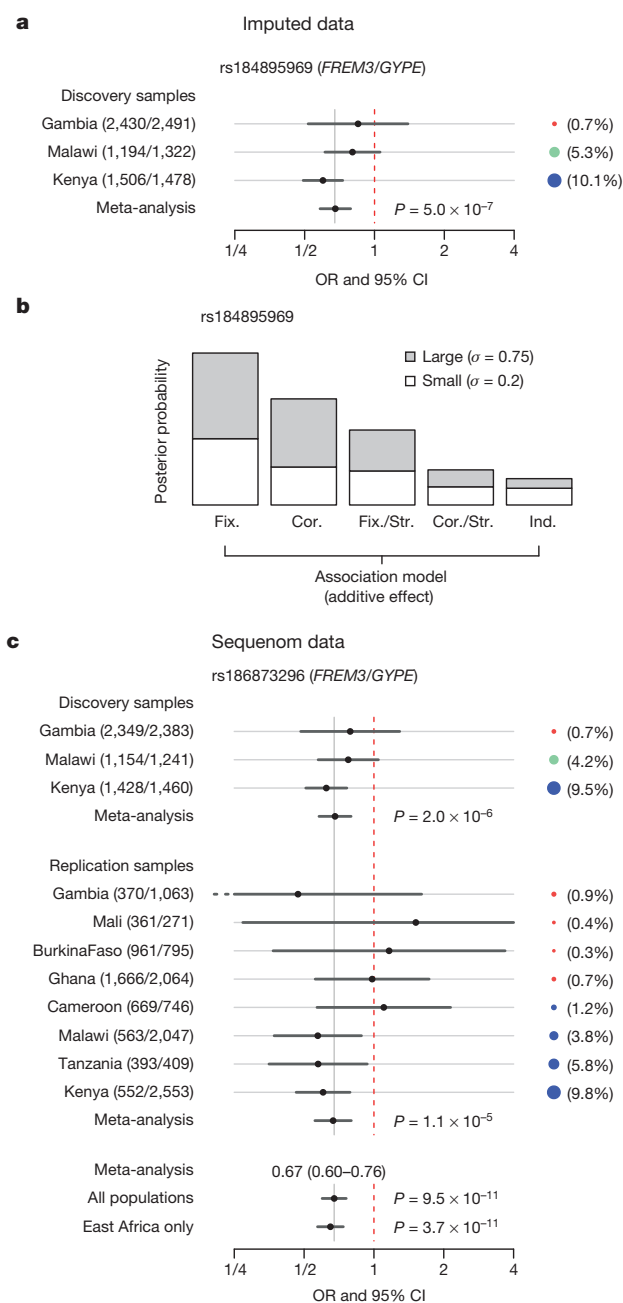


Figure 2 | Evidence for association at SNPs in the *FREM3/GYPE* region, assuming an additive model of association. **a**, Forest plot showing sample size, estimated OR and 95% CI for the lead imputed SNP in each population and under fixed-effect meta-analysis. The frequency of the protective allele in controls in each population is shown to the right. **b**, The posterior weight on different models of heterogeneity at rs184895969 under the prior used in the GWAS. Cor., correlated effects model; Cor./Str., correlated-structured-effect model; Fix., fixed effects model; Fix./Str., fixed-structured-effect model; Ind., independent effects model. Models are described in Methods. **c**, Forest plot for the Sequenom-typed SNP rs186873296 in discovery and replication samples, with fixed-effect meta-analysis across all populations and across East African populations (here taken as Kenya, Malawi, Tanzania and Cameroon.)

It has been proposed that *P. falciparum* was introduced into human populations from chimpanzees or gorillas in the recent past, but this remains a matter of intense debate^{25–27}. Population genetic data are consistent with an ancient origin followed by a marked expansion of the parasite population approximately 10,000 years ago, coincident with the introduction of agriculture²⁸. The *P. falciparum* genome

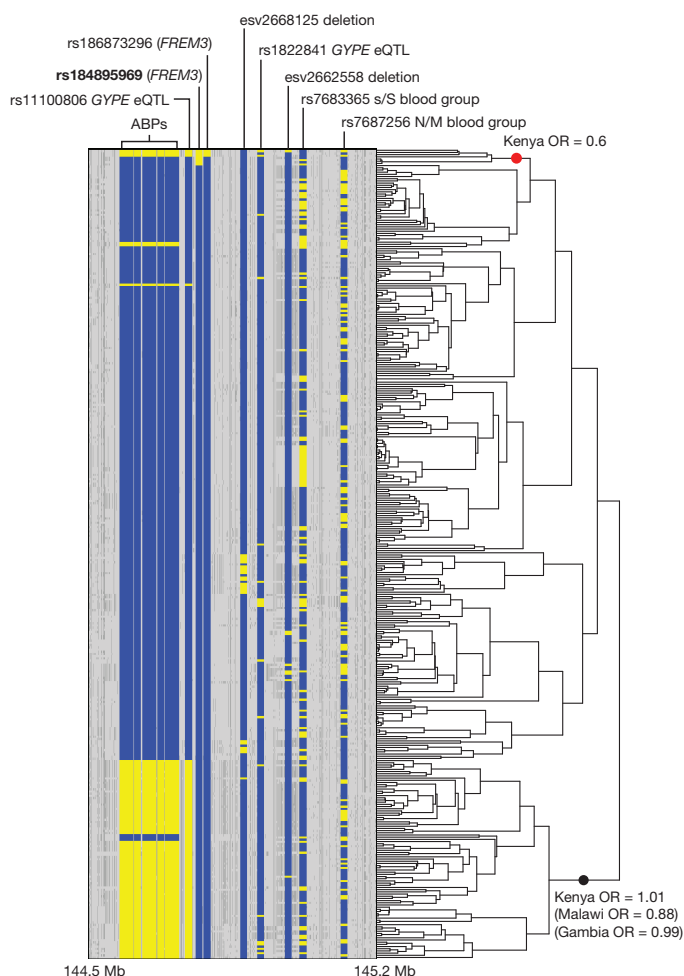


Figure 3 | Haplotype analysis across the *FREM3*/*GYPE* region. Left, haplotypes at 7,321 polymorphic SNPs between 144.5 Mb and 145.2 Mb on chromosome 4 in the LWK and YRI samples of the 1000 Genomes Project reference panel. Key variants (Methods and Supplementary Note 2) are enlarged for clarity and labelled, with reference and non-reference alleles coloured blue and yellow, respectively. Right, the estimated topology of the genealogical tree at rs184895969. Dots indicate the position of the inferred protective mutation in Kenya and the branch ancestral to the ABPs, and are labelled with the estimated OR.

possesses a huge repertoire of polymorphism⁸, and it is possible that the host and parasite genomes are engaged in a longstanding evolutionary arms race, each maintaining diversity to try to outflank the other²⁹. Intriguingly, the parasite surface receptor EBA-175, which directly binds glyophorin A during erythrocyte invasion, also contains structural polymorphisms with features of ancient dimorphism³⁰. The present findings provide new leads both to investigate these evolutionary mechanisms and to discover further genetic determinants of resistance to severe malaria in African children.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 19 September 2014; accepted 10 August 2015.

Published online 30 September 2015.

1. Malaria Genomic Epidemiology Network. Reappraisal of known malaria resistance loci in a large multicenter study. *Nature Genet.* **46**, 1197–1204 (2014).
2. Timmann, C. *et al.* Genome-wide association study indicates two novel resistance loci for severe malaria. *Nature* **489**, 443–446 (2012).
3. Band, G. *et al.* Imputation-based meta-analysis of severe malaria in three African populations. *PLoS Genet.* **9**, e1003509 (2013).
4. Leffler, E. M. *et al.* Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science* **339**, 1578–1582 (2013).

5. Fry, A. E. *et al.* Common variation in the ABO glycosyltransferase is associated with susceptibility to severe *Plasmodium falciparum* malaria. *Hum. Mol. Genet.* **17**, 567–576 (2008).
6. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
7. Teo, Y. Y., Small, K. S. & Kwiatkowski, D. P. Methodological challenges of genome-wide association analysis in Africa. *Nature Rev. Genet.* **11**, 149–160 (2010).
8. Manske, M. *et al.* Analysis of *Plasmodium falciparum* diversity in natural infections by deep sequencing. *Nature* **487**, 375–379 (2012).
9. Jallow, M. *et al.* Genome-wide and fine-resolution association analysis of malaria in West Africa. *Nature Genet.* **41**, 657–665 (2009).
10. Blumenfeld, O. O. & Huang, C. H. Molecular genetics of the glyophorin gene family, the antigens for MNSs blood groups: multiple gene rearrangements and modulation of splice site usage result in extensive diversification. *Hum. Mutat.* **6**, 199–209 (1995).
11. Sim, B. K., Chitnis, C. E., Wasniowska, K., Hadley, T. J. & Miller, L. H. Receptor and ligand domains for invasion of erythrocytes by *Plasmodium falciparum*. *Science* **264**, 1941–1944 (1994).
12. Mayer, D. C. *et al.* Glycophorin B is the erythrocyte receptor of *Plasmodium falciparum* erythrocyte-binding ligand, EBL-1. *Proc. Natl Acad. Sci. USA* **106**, 5348–5352 (2009).
13. Baum, J., Ward, R. H. & Conway, D. J. Natural selection on the erythrocyte surface. *Mol. Biol. Evol.* **19**, 223–229 (2002).
14. Ko, W. Y. *et al.* Effects of natural selection and gene conversion on the evolution of human glycophorins coding for MNS blood polymorphisms in malaria-endemic African populations. *Am. J. Hum. Genet.* **88**, 741–754 (2011).
15. Tarazona-Santos, E. *et al.* Population genetics of GYPB and association study between GYPB*s/s polymorphism and susceptibility to *P. falciparum* infection in the Brazilian Amazon. *PLoS ONE* **6**, e16123 (2011).
16. Wang, H. Y., Tang, H., Shen, C. K. & Wu, C. I. Rapidly evolving genes in human. I. The glycophorins and their possible role in evading malaria parasites. *Mol. Biol. Evol.* **20**, 1795–1804 (2003).
17. Stranger, B. E. *et al.* Patterns of *cis* regulatory variation in diverse human populations. *PLoS Genet.* **8**, e1002639 (2012).
18. Gurdasani, D. *et al.* The African Genome Variation Project shapes medical genetics in Africa. *Nature* **517**, 327–332 (2015).
19. Lanzillotti, R. & Coetzer, T. L. The 10 kDa domain of human erythrocyte protein 4.1 binds the *Plasmodium falciparum* EBA-181 protein. *Malar. J.* **5**, 100 (2006).
20. Ségurel, L. *et al.* The ABO blood group is a trans-species polymorphism in primates. *Proc. Natl Acad. Sci. USA* **109**, 18493–18498 (2012).
21. Su, Z., Cardin, N., The Wellcome Trust Case Control Consortium, Donnelly P. & Marchini, J. A Bayesian method for detecting and characterizing allelic heterogeneity and boosting signals in genome-wide association studies. *Stat. Sci.* **24**, 430–450 (2009).
22. Lee, P. H., O'Dushlaine, C., Thomas, B. & Purcell, S. M. INRICH: interval-based enrichment analysis for genome-wide association studies. *Bioinformatics* **28**, 1797–1799 (2012).
23. Karlsson, E. K., Kwiatkowski, D. P. & Sabeti, P. C. Natural selection and infectious disease in human populations. *Nature Rev. Genet.* **15**, 379–393 (2014).
24. Otto, T. D. *et al.* Genome sequencing of chimpanzee malaria parasites reveals possible pathways of adaptation to human hosts. *Nature Commun.* **5**, 4754 (2014).
25. Liu, W. *et al.* Origin of the human malaria parasite *Plasmodium falciparum* in gorillas. *Nature* **467**, 420–425 (2010).
26. Prugnolle, F. *et al.* A fresh look at the origin of *Plasmodium falciparum*, the most malignant malaria agent. *PLoS Pathog.* **7**, e1001283 (2011).
27. Rich, S. M. *et al.* The origin of malignant malaria. *Proc. Natl Acad. Sci. USA* **106**, 14902–14907 (2009).
28. Joy, D. A. *et al.* Early origin and recent expansion of *Plasmodium falciparum*. *Science* **300**, 318–321 (2003).
29. Gilbert, S. C. *et al.* Association of malaria parasite population structure, HLA, and immunological antagonism. *Science* **279**, 1173–1177 (1998).
30. Binks, R. H. *et al.* Population genetic analysis of the *Plasmodium falciparum* erythrocyte binding antigen-175 (*eba-175*) gene. *Mol. Biochem. Parasitol.* **114**, 63–70 (2001).

Supplementary Information is available in the online version of the paper.

Acknowledgements The Malaria Genomic Epidemiology Network Project is supported by the Wellcome Trust (WT077383/Z/05/Z) and the Bill & Melinda Gates Foundation through the Foundations of the National Institutes of Health (NIH; 566) as part of the Grand Challenges in Global Health Initiative. The Resource Centre for Genomic Epidemiology of Malaria is supported by the Wellcome Trust (090770/Z/09/Z). This research was supported by the Medical Research Council (MRC; G0600718; G0600230), the Wellcome Trust Biomedical ethics Enhancement Award (087285) and Strategic Award (096527). D.P.K. receives support from the MRC (G19/9). C.C.A.S. was supported by a Wellcome Trust Career Development Fellowship (097364/Z/11/Z). The Wellcome Trust also provides core awards to The Wellcome Trust Centre for Human Genetics (075491/Z/04; 090532/Z/09/Z) and the Wellcome Trust Sanger Institute (077012/Z/05/Z and 098051). The Mali MRTCC-BMP group is supported by a contract (N01AI85346) and a cooperative agreement (U19AI065683) from the National Institute of Allergy and Infectious Diseases (NIAID) and by a grant (D43TW001589) from the Fogarty International Centre, NIH to University of Maryland and University of Bamako and the Mali-NIAID/NIH International Center for Excellence in Research at the University of Sciences, Techniques, and Technology of Bamako. E.A. received partial funding from the European Community's Seventh Framework Programme (FP7/2007–2013) under grant agreement no. 242095–EVIMalaR and the

Central African Network for Tuberculosis, HIV/AIDS and Malaria (CANTAM) funded by the European and Developing Countries Clinical Trials Partnership (EDCTP). T.N.W. is funded by Senior Fellowships from the Wellcome Trust (076934/Z/05/Z and 091758/Z/10/Z) and through the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 242095-EVIMalaR. The KEMRI-Wellcome Trust Programme is funded through core support from the Wellcome Trust. C.M.N. is supported through a strategic award to the KEMRI-Wellcome Trust Programme by the Wellcome Trust (084538). Tanzania/Kilimanjaro Christian Medical College Joint Malaria Programme, Moshi, Tanzania received funding from MRC grant number G9901439. We acknowledge the work of B. Poudiougou and A. Niangaly for their help in collecting clinical data and biological samples for the Bamako study. We thank L. Jostins and M. Pirinen for advice on statistical analyses.

Author Contributions Details of author contributions are provided in the author list.

Author Information Genotype and phenotype data underlying this manuscript have been deposited in the European Genome-phenome Archive under accession number EGAS00001001311. Access to individual-level genotype data is available by application to an Independent Data Access Committee: see <http://www.malariagen.net/data>. For further details of data underlying this manuscript, see <http://www.malariagen.net/resource/14>. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.P.K. (dominic@sanger.ac.uk) or C.C.A.S. (chris.spencer@well.ox.ac.uk). A complete list of contributors to the MalariaGEN Consortium Project 1 has been published previously¹.

Malaria Genomic Epidemiology Network

Writing group Gavin Band¹, Kirk A. Rockett^{1,2}, Chris C. A. Spencer¹, Dominic P. Kwiatkowski^{1,2}; **Data analysis** Gavin Band¹, Quang Si Le¹, Geraldine M. Clarke¹, Katja Kivinen², Ellen M. Leffler¹, Kirk A. Rockett^{1,2}, Dominic P. Kwiatkowski^{1,2}, Chris C. A. Spencer¹; **Project management** Kirk A. Rockett^{1,2}, Chris C. A. Spencer¹, Victoria Cornelius¹, David J. Conway^{3,4}, Thomas N. Williams^{5,6}, Terrie Taylor^{7,8}, Dominic P. Kwiatkowski^{1,2}; **Study site lead investigators** David J. Conway^{3,4}, Kalifa A. Bojang³, Ogobara Doumbo⁹, Mahamadou A. Thera⁹, David Modiano¹⁰, Sodiomon B. Sirima¹¹, Michael D. Wilson¹², Kwadwo A. Koram¹², Tsiri Agbenyega^{13,14}, Eric Achidi¹⁵, Thomas N. Williams^{5,6}, Kevin Marsh⁵, Hugh Reyburn^{16,17}, Chris Drakeley^{16,17}, Eleanor Riley¹⁷, Terrie Taylor^{7,8}, Malcolm Molyneux¹⁸; **Clinical data and sample collection** Muminatou Jallow^{3,19}, Kalifa A. Bojang³, David J. Conway^{3,4}, Margaret Pinder³, Ogobara Doumbo⁹, Mahamadou A. Thera⁹, Ousmane B. Toure⁹, Salimata Konate⁹, Sibiri Sissoko⁹, Edith C. Bougouma¹¹, Valentina D. Mangano¹⁰, David Modiano¹⁰, Sodiomon B. Sirima¹¹, Lucas

N. Amenga-Etego²⁰, Anita K. Ghansah¹², Abraham V. O. Hodgson²⁰, Kwadwo A. Koram¹², Michael D. Wilson¹², Tsiri Agbenyega^{13,14}, Daniel Ansong^{13,14}, Anthony Enimil¹³, Jennifer Evans^{21,22}, Eric Achidi¹⁵, Tobias O. Apinijoh²³, Alexander Macharia⁵, Kevin Marsh⁵, Carolyn M. Ndila⁵, Charles Newton⁵, Norbert Peshu⁵, Sophie Uyoga⁵, Thomas N. Williams^{5,6}, Chris Drakeley^{16,17}, Alphaxard Manjurano^{16,17}, Hugh Reyburn^{16,17}, Eleanor Riley¹⁷, David Kachala¹⁸, Malcolm Molyneux¹⁸, Vysaul Nyirongo¹⁸, Terrie Taylor^{7,8}; **Sample processing, genotyping, data management and project coordination** Kirk A. Rockett^{1,2}, Katja Kivinen², Daniel Mead², Eleanor Drury², Sarah Auburn¹, Susana G. Campino², Bronwyn MacInnis², Jim Stalker², Emma Gray², Christina Hubbard¹, Anna E. Jeffreys¹, Kate Rowlands¹, Alieu Mendy¹, Rachel Craik¹, Kathryn Fitzpatrick¹, Sile Molloy¹, Lee Hart¹, Robert Hutton¹, Angeliki Kerasidou^{1,24}, Kimberly J. Johnson¹, Victoria Cornelius¹

¹Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK. ²The Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK. ³Medical Research Council Unit, Atlantic Boulevard, Fajara, PO Box 273, The Gambia. ⁴Department of Pathogen Molecular Biology, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK. ⁵KEMRI-Wellcome Trust Research Programme CGMRC, PO Box 230-80108, Kilifi, Kenya. ⁶Faculty of Medicine, Department of Medicine, Imperial College, Exhibition Road, London SW7 2AZ, UK. ⁷Blantyre Malaria Project, Queen Elizabeth Central Hospital, College of Medicine, P.O. Box 30096, Chichiri, Blantyre 3, Malawi. ⁸College of Osteopathic Medicine, Michigan State University, East Lansing, Michigan 48824, USA. ⁹Malaria Research and Training Centre, Faculty of Medicine University of Bamako, Bamako, Mali. ¹⁰University of Rome La Sapienza, Piazzale Aldo Moro 5, 00185 Rome, Italy. ¹¹Centre National de Recherche et de Formation sur le Paludisme (CNRFP), 01 BP 2208 Ouagadougou 01, Burkina Faso. ¹²Noguchi Memorial Institute for Medical Research, University of Ghana, P.O. Box LG 25, Legon, Accra, Ghana. ¹³Komfo Anokye Teaching Hospital, PO Box 1934, Kumasi, Ghana. ¹⁴Kwame Nkrumah University of Science and Technology, Kumasi, Ghana. ¹⁵Department of Medical Laboratory Sciences, University of Buea, P.O. Box 63, Buea, South West Region, Cameroon. ¹⁶Joint Malaria Programme, Kilimanjaro Christian Medical Centre, PO box 2228, Moshi, Tanzania. ¹⁷Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK. ¹⁸Malawi-Liverpool-Wellcome Trust Clinical Research Programme, Queen Elizabeth Central Hospital, College of Medicine, P.O. Box 30096, Chichiri, Blantyre 3, Malawi. ¹⁹Royal Victoria Teaching Hospital, Independence Drive, PO Box 1515, Banjul, The Gambia. ²⁰Navrongo Health Research Centre, PO Box 114, Navrongo, Ghana. ²¹Department of Molecular Medicine, Bernhard Nocht Institute for Tropical Medicine, Postfach 30 41 2, D-20324 Hamburg, Germany. ²²Kumasi Centre for Collaborative Research, School of Medical Sciences, KNUST, Kumasi, Ghana. ²³Department of Biochemistry & Molecular Biology, University of Buea, P.O. Box 63 Buea, South West Region, Cameroon. ²⁴The Ethox Centre, Nuffield Department of Population Health, University of Oxford, Old Road Campus, Oxford OX3 7LF, UK.

METHODS

Samples, ethics and clinical information. Samples were collected from nine partner projects from across sub-Saharan Africa (Extended Data Fig. 1a) as described previously^{1,3}. The studies and sample sets described in this manuscript form part of a larger ongoing project within the Malaria Genomic Epidemiology Network (MalariaGEN; <http://www.malariagen.net>). We used the World Health Organization definition of severe malaria, which comprises a broad spectrum of life-threatening clinical complications of *Plasmodium falciparum* infection^{29,31}. Investigators from study sites worked together to agree on principles for sharing data and standardizing clinical definitions, and to define best ethical practices across different local settings including the development of guidelines for informed consent. Relevant ethics committees are listed in Extended Data Fig. 1a. Further information on policies, research and the consent process may be found on the MalariaGEN website (<http://www.malariagen.net/community/ethics-governance>). No statistical methods were used to predetermine sample size.

DNA extraction and Sequenom typing. As described previously^{1,3}, all samples submitted to the MalariaGEN Resource Centre underwent a standard set of procedures that included quantification using picogreen, genotyping of 65 polymorphisms (including the polymorphism rs334 causing sickle haemoglobin, and three gender-typing SNPs) on the Sequenom iPLEX MassArray platform and matching to baseline clinical data (for example, gender, ethnic group and case–control status).

High-density genotyping. Three cohorts (Gambia, Malawi and Kenya) were genotyped on the Illumina HumanOmni2.5-4 (Kenya) and Illumina HumanOmni2.5-8 (Gambia, Malawi) platforms. As described previously³ we used three different calling algorithms (Illuminus³², Illumina's Gencall algorithm as provided in BeadStudio, and GenoSNP³³), each of which uses slightly different information in the data. We formed final genotype calls by taking consensus between the three algorithms. Genotypes where any two of the three calling algorithms were discordant, and genotypes where fewer than two algorithms were confident enough to make a call were treated as missing. This process showed improved calling, evaluated using Mendelian error counts in a subset of Kenyan samples, relative to each of the three algorithms separately (data not shown).

After genotype calling, we aligned genotypes to the forward strand of the human reference sequence (GRCh37), using both the Illumina-supplied manifest and publicly available strand files (<http://www.well.ox.ac.uk/~wrayner/strand>) obtained by mapping allele probes to the reference by BLAT³⁴. We removed SNPs whose position or strand mismatched between the Illumina manifest and the strand file. To simplify analyses, we restricted attention to the set of SNPs having the same name, chromosome, position, strand, and probe sequences across the two genotyping platforms. Because the Omni2.5M contains multiple probes for some variants, we further removed SNPs to ensure positions were unique. In total we were left with 2,322,985 SNPs in each cohort across the autosomes and the X and Y chromosomes. We note that SNPs annotated as lying on the pseudo-autosomal region (PAR) of the X chromosome were not included, as these had position equal to 0 in the manifest for the HumanOmni2.5-4 array. Finally, we flipped alleles where necessary so that in downstream analyses the first allele always corresponds to the reference allele of the human genome sequence.

Sample quality control. We performed sample quality control separately on each cohort by computing autosome-wide averages of normalized X channel intensity, normalized Y channel intensity, and heterozygosity and missingness based on the consensus call. To identify outlying samples we applied ABERRANT³⁵, adjusting the λ parameter per cohort to account for differences in genotyping quality (Extended Data Fig. 1d–f). In Gambia, a tail of samples showing low heterozygosity but otherwise appearing to be well typed was apparent. We explicitly included these samples in downstream analyses.

To estimate genome-wide relatedness between samples, we selected a list of 178,775 high-quality SNPs satisfying the criteria missingness <1%, minor allele frequency (MAF) >1% and thinned to be at least 0.005 cM apart and to have pairwise $r^2 < 0.3$. Treating each cohort separately, we used SHELLFISH (<http://www.stats.ox.ac.uk/~davison/software/shellfish/shellfish.php>) to compute a matrix of pairwise relatedness values $R = (r_{ij})$ between samples, where r_{ij} denotes the genome-wide average covariance of frequency-normalized genotypes in individuals i and j . In samples with few close relationships, the value of r_{ij} can be thought of as an estimate of kinship³⁶ with, for example, values close to 1 representing identity between samples, and values close to zero reflecting a lack of close relatedness (relative to the rest of the sample). To remove samples with duplicate typing and close relationships, we excluded one of each pair of samples with $r_{ij} > 0.2$, taking all remaining samples through to phasing and imputation. Extended Data Figure 1c lists the number of samples before and after quality control, and the number removed by each quality control step.

We used SHELLFISH to compute principal components on the post-quality-control sample set. Consistent with our removal of poor quality, duplicate and

closely related samples, principal components plots show no substantially outlying samples (see Extended Data Fig. 4a–c). The top few principal components in each cohort reflect substantial population structure, as evidenced by colouring by reported ethnicity. As found previously^{3,9}, the top principal components are also significantly correlated with case/control status (Extended Data Fig. 4d), indicating that population structure may act as a confounding factor in association analyses if not controlled.

Genotyping quality control. Treating each cohort separately, we used SNPTEST (http://mathgen.stats.ox.ac.uk/genetics_software/snpTest/snpTest.html) to test for association at each autosomal SNP using a genotypic model of association that allows for different effects at heterozygote and homozygote genotypes, including the leading five principal components as covariates. Genotypic model association tests are particularly sensitive to confounding by genotyping error³⁷, and we used the resulting P values as a guide to finding appropriate quality control criteria. To detect potential spurious genotypes due to batch effects, we modelled genotypes as predicted by the leading five principal components and case/control status in a linear regression framework in R³⁸, and tested whether including an indicator of the plate on which each sample was genotyped contributed significantly to model fit. We refer to this as the 'plate test'. For downstream analyses we excluded SNPs with MAF <1%, missing data proportion >5% (in Gambia and Malawi) or >2.5% (in Kenya, which had fewer missing genotypes overall), Hardy–Weinberg $P < 1 \times 10^{-20}$ in controls, or plate-test $P < 0.01$. We inspected cluster plots of all remaining SNPs showing association test $P < 1 \times 10^{-5}$ (Gambia) or $P < 1 \times 10^{-4}$ (Kenya, Malawi) and excluded those with clear genotyping problems. Extended Data Figure 2a shows the number of SNPs excluded by each criterion. The post-exclusion genome-wide association analysis is shown in Extended Data Fig. 2c.

SNP quality control on the X chromosome was performed as on the autosomes, with a few differences as follows. We treated males and females separately, using a genotypic model of association in females and an allelic model in males (who have only one copy of the X chromosome). Because genotype calling was performed blind to the gender of samples, some males appear to be heterozygous at some X chromosome SNPs. We treated all such heterozygous calls as missing. We computed missing data proportion and plate-test P values in males and females separately, and tested for departure from Hardy–Weinberg in female controls and for differences in frequency between males and females. We excluded SNPs with MAF <1%, missingness >5% (Gambia, Malawi) or 2.5% (Kenya) in males or females, or plate-test $P < 0.01$ in males or females. We further excluded SNPs with Hardy–Weinberg $P < 1 \times 10^{-20}$ in female controls or showing significant difference in allele frequencies between males and females ($P < 1 \times 10^{-20}$). Extended Data Figure 2b shows the number of SNPs excluded by each criterion.

Some regions of the X chromosome showed an elevated number of male heterozygote calls, contributing to the high number of SNPs excluded due to missingness in males. These included SNPs in the pseudo-autosomal regions at either end of the chromosome³⁹ (indicating that XY designation in the chip manifests does not adequately cover these regions) as well as SNPs within the X transposed region near the centromere.

Phasing and imputation. We phased genotype data in each cohort separately using SHAPEIT v2.864 (ref. 40), specifying 200 hidden states and an effective population size of 17,469, as recommended for African populations by the SHAPEIT documentation, and phasing each chromosome separately. We used IMPUTE v2.3.0 (refs 41, 42) to impute phased genotypes into the 1000 Genomes⁶ Phase I integrated variant set (version of 24th August 2012, as downloaded from the IMPUTE website, which we refer to here as the 1000 Genomes reference panel) in 5Mb chunks with a buffer region of 500kb. For phasing and imputation we used the combined HapMap recombination map in build 37 coordinates included with the 1000 Genomes reference panel. Unless otherwise stated, downstream analyses included only SNPs with MAF >0.5% and IMPUTE info measure >0.75.

Assessment of imputation. At each genotyped SNP, IMPUTE computes squared correlation (referred to here as accuracy) and concordance between typed genotypes and genotypes obtained by masking the SNP and re-imputing. To assess imputation performance, we plotted the cumulative distribution of concordance (Extended Data Fig. 3a) and accuracy (Extended Data Fig. 3b), accuracy by frequency (Extended Data Fig. 3d) as well as average per-sample accuracy (Extended Data Fig. 3c). We also assessed accuracy relative to direct typing on the Sequenom platform at variants typed in the *FREM3/GYPE*, *ARL14* and *INPP4B* regions of association as described later.

Association testing. We used SNPTEST to test for association at approximately 38 million variants obtained through imputation, including five principal components as covariates to control for population structure separately in each cohort. SNPTEST uses a missing data likelihood to account for the uncertainty in genotypes at imputed SNPs. We fit additive, dominant, recessive, and heterozygote models of association and ran SNPTEST separately in each imputation chunk.

Below, we refer to the estimated effect size for population i and mode of inheritance m as $\beta_{i,m}$ and its estimated standard error as $SE_{i,m}$.

Frequentist meta-analysis. For each SNP and each mode of inheritance (additive, dominant, recessive, heterozygote) we computed the fixed-effect inverse variance-weighted meta-analysis effect size, standard error, and P value. In this context, fixed-effect meta-analysis assumes a single true effect size that is identical between the three cohorts, and can be thought of as finding the maximum likelihood estimate of the effect size under the assumption that the likelihood in cohort i is proportional to the density of the normal distribution with mean $\beta_{i,m}$ and standard error $SE_{i,m}$.

Bayesian meta-analysis. We have previously^{1,3} used Bayesian meta-analysis techniques to allow for between-population heterogeneity of effect sizes in these three cohorts. Here, we applied this method to compute Bayes factors for association under four modes of inheritance and six different models of correlation between cohorts. In this framework, true effect sizes are modelled by a multivariate normal distribution centred on zero and with a given prior covariance matrix that can be written as $\sigma^2 P$, where σ^2 is a prior variance controlling the magnitude of plausible effects and P is a prior correlation matrix.

We used the following correlation models.

Fixed effects (all elements of P equal to 1): as with frequentist fixed-effect meta-analysis, this assumes effect sizes are equal across the three cohorts.

Correlated effects (all off-diagonal elements of P equal to 0.96): this assumes effect sizes are similar but allows for some variability.

Independent effects (all off-diagonal elements of P equal to 0.1): this assumes there is little similarity between effect sizes in different cohorts.

Structured effects: we also considered models where effects in the two East African populations (Kenya and Malawi) were more similar to each other than to that in The Gambia. We assumed either effects were fixed between Kenya and Malawi and correlated ($\rho = 0.96$) between East Africa and The Gambia (we refer to this as the fixed-structured-effect model), or that effects were correlated ($\rho = 0.96$) between Kenya and Malawi and largely independent with The Gambia ($\rho = 0.1$) (referred to as the correlated-structured-effect model).

We used prior variance parameters of $\sigma^2 = 0.2^2$, reflecting a belief in relatively small effects (OR < 1.5 with 95% probability)³⁷, and $\sigma^2 = 0.75^2$, reflecting a belief in larger effects (OR < 4.5 with 95% probability).

To form a single summary measure of evidence for association we formed a model-averaged Bayes factor (BF), referred to as the mean BF and denoted BF_{avg} , as a weighted average of model-specific Bayes factors using the following weighting scheme. We assigned weights of 0.4 for additive mode of inheritance and 0.2 for dominant, recessive, or heterozygote modes of inheritance, reflecting a belief that variants which tag causal variants by linkage disequilibrium are more likely to display additive effects. We assigned a weight of 0.4, 0.2, and 0.1 to fixed-, correlated-, and independent-effect models, respectively, and 0.2 and 0.1 to fixed-structured and correlated-structured effects models. Finally we assigned weight of 0.5 to small-effect models ($\sigma^2 = 0.2^2$) and large-effect models ($\sigma^2 = 0.75^2$). Overall prior weights were assigned by multiplying across these categories; for example, the model representing small effect size distribution, fixed-effect across cohorts with additive mode of inheritance was assigned prior weight equal to $0.5 \times 0.4 \times 0.4 = 0.08$, while the model representing small, correlated-structured, dominant effects was assigned a prior weight of $0.5 \times 0.1 \times 0.2 = 0.01$. For each SNP we also recorded the model having the highest posterior weight, and refer to its Bayes factor as the maximum BF (denoted BF_{max}). Extended Data Figure 5a depicts slices through the combined prior on effect sizes across three cohorts for additive effect models. The mean BF behaves similarly to a minimum over all four fixed-effect meta-analysis P values, but additionally captures effects that vary between cohorts (Extended Data Fig. 5b).

As described previously^{1,3}, to compute model-specific Bayes factors efficiently we used an approximation of the likelihood by the density of a normal distribution with the estimated mean ($\beta_{i,m}$) and standard error ($SE_{i,m}$) in each cohort. Thus, overall, observed effect sizes are modelled as normally distributed around zero with covariance that depends on the prior covariance in true effect sizes and on model standard errors

$$(\beta_{i,m}) \sim N(0, \sigma^2 P + V)$$

where V is a diagonal matrix with i th diagonal entry equal to the squared standard error, $SE_{i,m}^2$. The approximate or asymptotic Bayes factor can then be computed by evaluating a ratio of two normal densities. In the univariate case this method is the same as described previously⁴³. To facilitate working with genome-wide meta-analysis results, we wrote custom software to compute frequentist and Bayesian meta-analyses, and stored details of genotype counts, association model fit, and meta-analysis results directly in a SQLite database file.

Further discussion of the Bayesian approach can be found in Supplementary Note 4.

X chromosome association testing and meta-analysis. Association testing on the X chromosome was performed as for the autosomes with a few differences as follows. We ran SNPTEST separately in males and females, estimating effect sizes under additive, dominant, recessive and heterozygote modes of inheritance in females. In this usage, SNPTEST assumes a model of complete inactivation so encodes male genotypes as 0/1 and females as 0/0.5/1 for an additive mode of inheritance. We then meta-analysed the six gender-specific association analyses to produce frequentist fixed-effect and model-averaged Bayesian meta-analyses.

For Bayesian analysis, in addition to summing over models of between-population heterogeneity, we adopted the view that differences in sex might lead to heterogeneity in effect. We therefore included models of heterogeneity between males and females as follows. Let ρ_{sex} denote the correlation between effects in male and female samples within a single population. We included models where males and females have the same effect ($\rho_{sex} = 1$, termed fixed-sex model and given prior weight 0.45), correlated effects ($\rho_{sex} = 0.96$, termed correlated-sex model and given prior weight 0.225) or essentially independent effects ($\rho_{sex} = 0.1$, termed independent-sex model and given prior weight 0.225). Because some parts of the X chromosome escape inactivation^{39,44}, we also included a model where the effect in females is twice that in males (given prior weight 0.1).

To fully specify prior correlation for each model, for each pair of populations A , B we also need to specify the correlation (denoted ρ_{cross}) between males in A and females in B . We set $\rho_{cross} = \rho_{pop} \times \rho_{sex}$ where ρ_{pop} is the chosen prior correlation between same-sex samples in A and B (that is, $\rho_{pop} = 1, 0.96$ or 0.1 as defined earlier).

As earlier, we formed overall weights by multiplying across categories, so that, for example, the model of small, additive effects that are fixed across populations and across sexes had prior weight $0.5 \times 0.4 \times 0.4 \times 0.45 = 0.036$.

Linear mixed model analysis. To compare association test results for logistic regression as implemented in SNPTEST with the use of a linear mixed model, we reran association test scans in each discovery cohort using the program MMM⁴⁵. We used the same genome-wide relatedness matrix as used to compute principal components (see above) and assumed an additive model of association. We plotted the $-\log_{10}(P \text{ value})$ for MMM against the corresponding $-\log_{10}(P \text{ value})$ based on the SNPTEST scan (which used five principal components as described above), for each discovery population and for fixed-effect meta-analysis (Extended Data Fig. 4f). P values under both methods at tier 1 loci are listed in Supplementary Table 1.

Lead SNPs, region and tier definitions. We formed a list of lead SNPs within approximately independent regions of interest as follows. We restricted to variants with IMPUTE info measure at least 0.75 across all three populations and ranked variants by the model-averaged Bayes factor (highest to lowest). We iteratively picked lead SNPs from the top of the list and excluded other variants within a recombination interval of $0.25 \text{ cM} \pm 25 \text{ kb}$ centred at the lead SNP (referred to later as the association region), continuing until no more SNPs remained with $BF_{avg} > 250$ or $BF_{max} > 2,500$.

We grouped lead SNPs into two tiers as follows: tier 1 containing all lead SNPs with $BF_{avg} > 2,500$ or $BF_{max} > 25,000$, and tier 2 containing all lead SNPs not in tier 1 with $BF_{avg} > 1,000$ or $BF_{max} > 10,000$.

In total, across the autosomes and the X chromosome there were 34 regions in tier 1 and 73 in tier 2, with association regions covering approximately 13 Mb and 26 Mb of the genome, respectively.

Regional association analysis. For each tier 1 and 2 region, we examined the pattern of association in the region, generating a regional association plot for the region annotated with details of the meta-analysis for the lead SNP as follows.

In each region we computed pairwise linkage disequilibrium statistics between the lead SNP and surrounding SNPs using best-guess imputed haplotypes for control samples across three populations. To facilitate this computation, we stored imputed haplotypes in a SQLite-format database allowing us effective random access to haplotype data. We used R to compute both Pearson correlation coefficient (r^2) and Lewontin's $|D'|$ between the lead SNP and all other SNPs in the region.

For each SNP in tier 1 and 2 we plotted $\log_{10}(BF_{avg})$ in the association region around the lead SNP plus 1 Mb on either side, colouring points according to linkage disequilibrium, with outer circles representing r^2 and inner circles representing $|D'|$. We further annotated SNPs that were typed in at least one cohort (using black plus signs) and SNPs that had Sequenom genotype data available (with black triangles).

We plotted all ABPs (supplementary tables 4 and 5 from ref. 4), expression quantitative trait loci (eQTLs) from the Genotype-Tissue Expression project⁴⁶ (GTEx), and previously reported⁴⁷ GWAS loci. Where SNPs matched an imputed or typed variant, we computed linkage disequilibrium statistics and coloured these points as described earlier. We plotted all RefSeq genes (as downloaded from the UCSC Genome Browser MySQL database⁴⁸ on 18 March 2013) in the region,

annotating the direction of transcription. We further plotted local recombination rate estimates from the HapMap combined recombination map included with the IMPUTE haplotypes described earlier.

We annotated plots with effect sizes and confidence intervals for the lead SNP for each mode of inheritance (additive, dominant, recessive and heterozygote) that informed the model averaging. We also produced barplots showing the posterior distribution on models of association using the prior weights described earlier. Finally, for each region, we produced and inspected cluster plots for all typed SNPs with $BF_{avg} \geq 10$ that were not excluded by quality control. As phasing fills in missing genotypes and potentially improves genotype calling based on linkage disequilibrium with surrounding SNPs, we coloured plots based on genotype calls taken from the output of phasing.

Website. Regional association plots and cluster plots can be viewed online at <http://www.malariagen.net/resource/14>.

Validation and replication typing. On the basis of a preliminary version of the data presented here we selected SNPs for typing on the Sequenom platform across the whole MalariaGEN Consortium Project 1 sample set, which includes the discovery samples, further cases and controls in the same populations that were not included in the GWAS, and further large sample sets from five other populations from sub-Saharan Africa (see Extended Data Fig. 1a). Data were available for SNPs tagging the lead markers in the *FREM3/GYPE*, *INPP4B* and *ARL14* regions ($r^2 > 0.5$ in controls, as estimated using the EM algorithm) as well as other regions not represented in tier 1.

Replication analysis. Replication analysis using Sequenom data was restricted to the set of samples with less than 10% missingness as measured across the set of 70 SNPs chosen for replication typing. In each population we conducted logistic regression in R, including five principal components (discovery samples) or reported ethnicity (replication samples) as covariates to control for population structure. For each GWAS lead SNP, we examined each Sequenom SNP in the region and computed r^2 and $|D'|$ with the lead SNP. To allow for the effects of incomplete linkage disequilibrium on the Bayesian model fitting, we recomputed the Bayesian analysis in the discovery samples based on Sequenom genotypes at each SNP, to obtain a Sequenom-based mean Bayes factor and a 'best model' Bayes factor at the model with highest posterior weight. Where linkage disequilibrium is incomplete or where imputation is imperfect, this model may differ from the best model for imputed data.

For each SNP we computed fixed-effect meta-analysis across discovery samples, across replication samples, and across all samples. We also computed the Bayes factor for replication samples ($BF_{replication}$) for the model with highest posterior weight in the discovery samples. To compute an overall Bayes factor for association, we combined the discovery BF_{avg} computed at the imputed lead marker with the replication Bayes factor at the Sequenom SNP, as $BF_{overall} = BF_{avg} \times BF_{replication}$. We use the imputed lead marker here because the number of discovery samples directly typed was smaller than the number of imputed samples. Conditional on the lead imputed and replication markers reflecting the same signal of association, this $BF_{overall}$ represents an overall measure of the evidence for association at the locus that reflects all the samples in our study.

A discussion of the replication evidence in the *FREM3/GYPE*, *INPP4B* and *ARL14* regions can be found in Supplementary Note 1.

MalariaGEN encourages individual study sites to perform more detailed analyses of local patterns of disease association, and a Tanzanian-focused analysis of *FREM3* and other candidate SNPs that were genotyped as part of this study is reported elsewhere⁴⁹.

Credible interval analysis. In a given region of the genome, under the assumption that exactly one variant (that is accessible to our typing or imputation) is causal, the posterior probability that each variant is the causal variant can be computed by a simple reweighting of Bayes factors⁵⁰. Using this approach we computed 95% and 99% credible intervals (that is, the smallest set of SNPs accounting for 95% or 99% of the posterior mass) for variants in the association region around rs184895969, plus a margin of 50 kb at either end (Extended Data Fig. 7b). We noted that rs181620317, which is annotated as a missense mutation for the gene *FREM3*, is within the 95% confidence interval in our data. We note that while this analysis is simple and appealing, its interpretation depends on assumptions about the true disease model, and on the behaviour of imputation in the region⁵⁰; in particular, the difficulty of imputing variants around the three glycoporphin genes (Extended Data Fig. 8) might make this analysis fail to capture putatively important variation within or around *GYPB*, *GYPB*, or *GYPE*.

Sequence homology and alignability in the glycoporphin region. To investigate the location of our GWAS signal with respect to the pattern of sequence homology around the three glycoporphin genes, we generated a dot plot (Extended Data Fig. 8a) showing co-occurrence of short segments of DNA (k -mers; $k = 25$ or 100 bases) in the human reference sequence⁵¹ in the region. Considerable sequence

homology was observed over a region of about 350 kb covering the three genes. Our lead GWAS marker and the ABPs lie just outside this region.

The high level of homology should affect our ability to align probes or sequences to this part of the genome. To confirm this, we also plotted the UCSC alignability track ('CRG alignability 100'; Extended Data Fig. 8c), which shows the degree to which the 100-mer starting at each position in the reference sequence is alignable (with a value of $1/n$ indicating that the 100-mer aligns to n positions across the genome, allowing up to two mismatches). As expected, even short regions of shared sequence affect alignability considerably, with alignability dropping to an average of about 0.7 within the large region of homology. Similarly, we plotted imputation performance (as measured by the IMPUTE info measure) and observed a marked drop within the region of sequence homology.

Structural variation in the glycoporphin region. We attempted to identify structural variation in the glycoporphin region by examining sequence read data generated by the 1000 Genomes Project, using the set of BAM files available from the 1000 Genomes Project in October 2014, downloaded from <ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data>. These data contain a mixture of read lengths, with most reads of at least 90 bp and some reads of 76 bp. We considered only reads with mapping quality at least 20. For each sample we computed coverage across the glycoporphin region, normalized by average coverage computed across chromosome 1, and refer to these values here as 'coverage'.

Coverage values across the region were correlated with genome alignability (as defined earlier), with coverage dropping substantially in regions of low alignability. We therefore restricted attention to the set of perfectly alignable positions, here defined as positions such that every 100-mer overlapping the position in the reference sequence aligns uniquely, allowing up to two mismatches (and computed using the CRG alignability 100 track described earlier). As expected, coverage at perfectly alignable positions showed little or none of the variation present in genome alignability (Extended Data Fig. 8e–g).

Two large structural variants, present at frequencies of at least 1% in LWK + YRI, were evident in coverage data. Genotypes for both these variants were called by the 1000 Genomes Phase I and are referred to as esv2668125 (also called MERGED_DEL_2_26708) and esv2662558 (also called MERGED_DEL_2_26722). Both deletions putatively represent deletions of all or part of *GYPB*. We noted some uncertainty as to the location of the breakpoints of both deletions—with breakpoints identified by the 1000 Genomes Project differing from the breakpoints as they appeared in coverage data by at least 10 kb. We further noted that three samples (NA18519, NA19185, NA19222) that were called as heterozygote for one or both deletions by the 1000 Genomes Project appeared to have homozygous genotypes (Extended Data Fig. 8e–g, red blue and green lines).

eQTL analysis. To investigate the effect of associated SNPs in the *FREM3* region on expression of nearby genes, we downloaded publicly available data on RNA expression levels¹⁷ in HapMap samples, and plotted expression levels of genes in the region (Fig. 1) against genotypes at the lead SNP and other SNPs of interest in African samples (YRI + LWK, Extended Data Fig. 9). Data was available for genes *INPP4B*, *USP38*, *GAB1*, *SMARCA5*, *GYPE*, *GYPB*, *GYPB*, *GYPB*, *HHIP*, *ANAPC10* and *ABCE1*, with three probes available for *GYPE*, two for *GAB1* and one probe for each of the other genes. Five samples in this data set (NA19318, NA19324, NA19377, NA19429, NA19190) carry the protective allele at the lead marker rs184895969, while only two carry the protective allele at the directly typed rs186873296. For each gene and SNP we tested for a trend of genotype on expression using linear regression. Extended Data Fig. 9 shows all probes for the glycoporphins as well as other regional genes for which a P value less than 0.05 was observed.

GENECLUSTER analysis. We used the program TREESIM²¹ to estimate an ancestral recombination graph for the LWK and YRI individuals in the 1000 Genomes Project haplotype data in a region from 144.6 Mb to 144.8 Mb on chromosome 4 centred on the lead SNP in the *FREM3/GYPE* region. We then ran GENECLUSTER²¹ in the region, allowing it to assign either one or two causal mutations in each marginal genealogy. GENECLUSTER works by probabilistically assigning study individuals to the tips of the tree estimated by TREESIM, and attempts to explain case/control status by assigning causal mutations to the branches of the tree in a Bayesian framework. This analysis is somewhat similar in spirit to our marginal SNP analysis, but has important differences. In particular, GENECLUSTER may detect associated mutations anywhere on the tree (which may not correspond directly to any typed or imputed variant), and can assess models of association involving more than a single causal variant. However, GENECLUSTER does not take into account principal components or otherwise control for population structure, and for computational reasons we only included the Luhyan (LWK) and Yoruban (YRI) populations from the 1000 Genomes reference panel in the analysis.

To investigate the relationship between ancestry and variants of interest, we plotted IMPUTE reference panel haplotypes in the region ordered by the marginal tree at the position of the imputed lead marker (rs184895969), annotating the lead

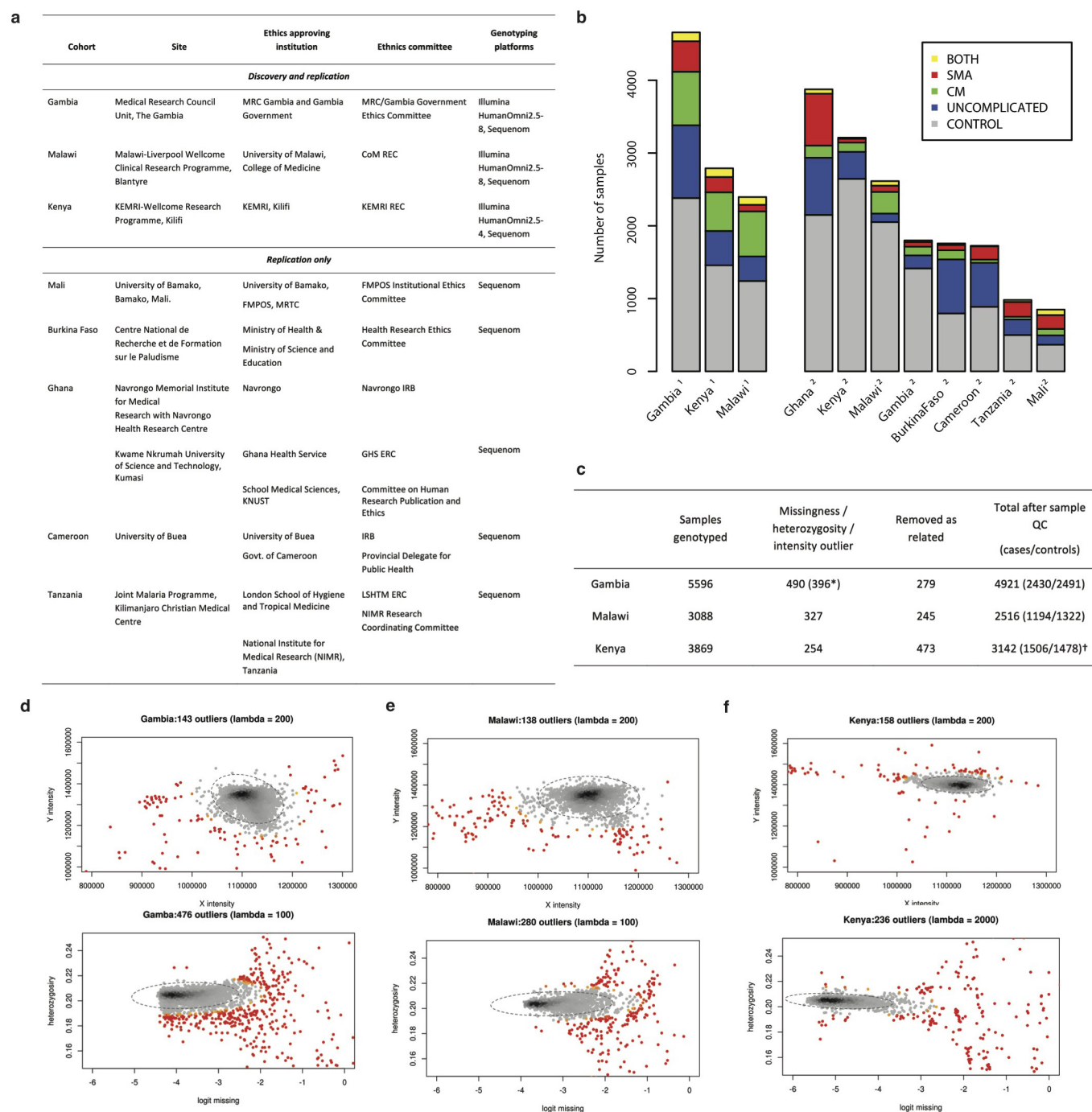
imputed and Sequenom-typed markers, ABPs, previously reported *GYPE* eQTLs⁵², common deletions, and variants determining the M/N and S/s blood groups as described in Supplementary Note 2 (Fig. 3).

Analysis of enrichment of malaria-associated loci in functional categories. Full details of enrichment analyses are provided in Supplementary Note 3.

Code availability. Executables and source code for *inthin* are available at <http://www.well.ox.ac.uk/~gav/inthin>.

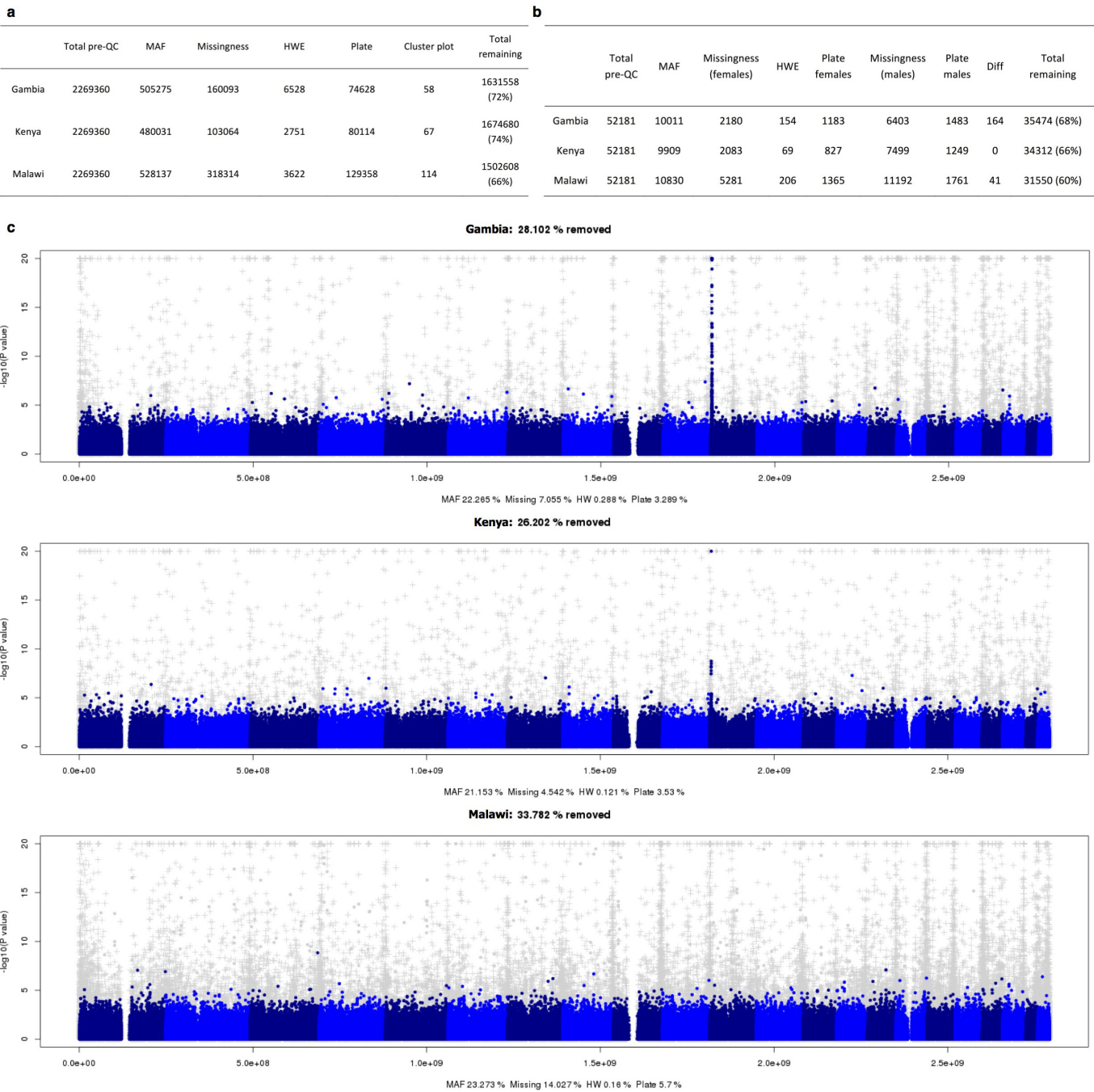
Further software for generating key results for this paper will be made available at <http://www.malariagen.net/resource/14>.

31. World Health Organization. Severe falciparum malaria. *Trans. R. Soc. Trop. Med. Hyg.* **94** (suppl. 1), S1–S90 (2000).
32. Teo, Y. Y. Genotype calling for the Illumina platform. *Methods Mol. Biol.* **850**, 525–538 (2012).
33. Giannoulatou, E., Yau, C., Colella, S., Ragoussis, J. & Holmes, C. C. GenoSNP: a variational Bayes within-sample SNP genotyping algorithm that does not require a reference population. *Bioinformatics* **24**, 2209–2214 (2008).
34. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
35. Bellenguez, C. *et al.* A robust clustering algorithm for identifying problematic samples in genome-wide association studies. *Bioinformatics* **28**, 134–135 (2012).
36. Astle, W. & Balding, D. J. Population structure and cryptic relatedness in genetic association studies. *Stat. Sci.* **24**, 451–471 (2009).
37. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
38. R Development Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2015).
39. Ross, M. T. *et al.* The DNA sequence of the human X chromosome. *Nature* **434**, 325–337 (2005).
40. Delaneau, O., Zagury, J. F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods* **10**, 5–6 (2013).
41. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genet.* **44**, 955–959 (2012).
42. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).
43. Wakefield, J. Bayes factors for genome-wide association studies: comparison with *P*-values. *Genet. Epidemiol.* **33**, 79–86 (2009).
44. Carrel, L., Cottle, A. A., Goglin, K. C. & Willard, H. F. A first-generation X-inactivation profile of the human X chromosome. *Proc. Natl Acad. Sci. USA* **96**, 14440–14444 (1999).
45. Pirinen, M., Donnelly, P. & Spencer, C. C. A. Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. *Ann. Appl. Stat.* **7**, 369–390 (2013).
46. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
47. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
48. Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32**, D493–D496 (2004).
49. Manjurano, A. *et al.* *USP38*, *FREM3*, *SDC1*, *DDC*, and *LOC727982* gene polymorphisms and differential susceptibility to severe malaria in Tanzania. *J. Infect. Dis.* **212**, 1129–1139 (2015).
50. The Wellcome Trust Case Control Consortium *et al.* Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nature Genet.* **44**, 1294–1301 (2012).
51. Hillier, L. W. *et al.* Generation and annotation of the DNA sequences of human chromosomes 2 and 4. *Nature* **434**, 724–731 (2005).
52. Zeller, T. *et al.* Genetics and beyond—the transcriptome of human monocytes and disease susceptibility. *PLoS One* **5**, e10693 (2010).



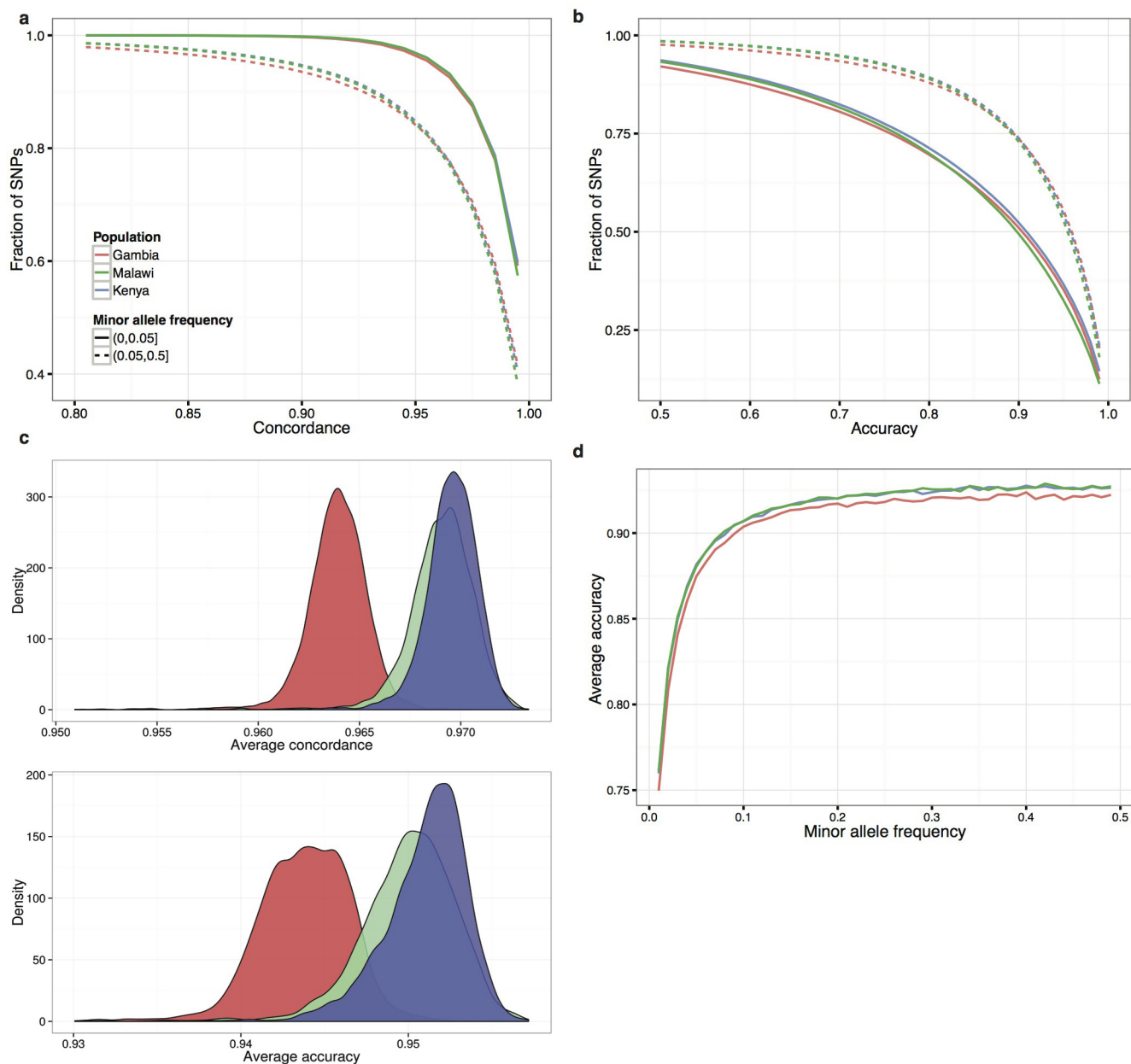
Extended Data Figure 1 | Sample collections included in the study. **a**, Study sites and ethics approving institutions. **b**, Phenotypic makeup of discovery and replication samples from each site. ‘Uncomplicated’ refers to case individuals who were not identified as cerebral malaria (CM) or severe malarial anaemia (SMA) cases. ‘Both’ refers to individuals who have both CM and SMA phenotypes. **c**, Overall sample counts and number of samples excluded by

each quality control (QC) criterion. Asterisk denotes the number of samples removed after explicitly including samples with low heterozygosity in Gambia. Dagger: the Kenyan cohort included parents of a subset of case samples; these were not used in subsequent analyses **d–f**. Plots of average genome-wide heterozygosity and missingness with outliers coloured, as outputted by the ABERRANT algorithm.



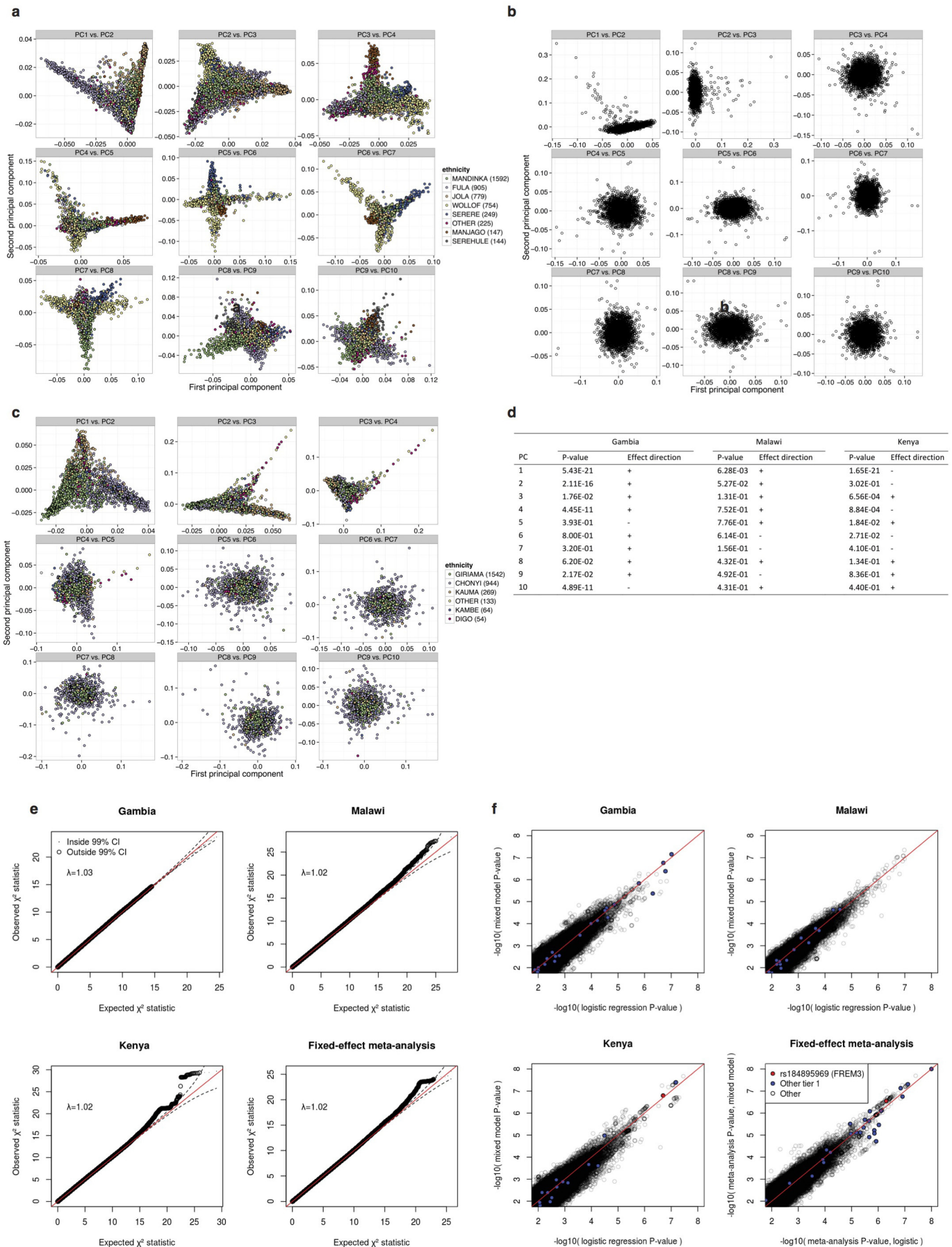
Extended Data Figure 2 | Genotyped SNP quality control for the three discovery cohorts. **a, b,** Total numbers of pre- and post-quality control SNPs on the autosomes (**a**) and the X chromosome (**b**), and numbers of SNPs excluded by each quality control (QC) criterion. Diff, the test of difference in frequency between males and females; HWE, Hardy–Weinberg equilibrium; Plate, plate test of association. Details of quality control are given in

Methods. **c,** Plot showing the $-\log_{10}(P \text{ values})$ for the genotypic association test in the discovery data including the first five principal components as covariates. Grey dots show SNPs that are removed due to the quality control as defined in Methods. The total fraction of SNPs removed from each cohort is given at the top of the plot.



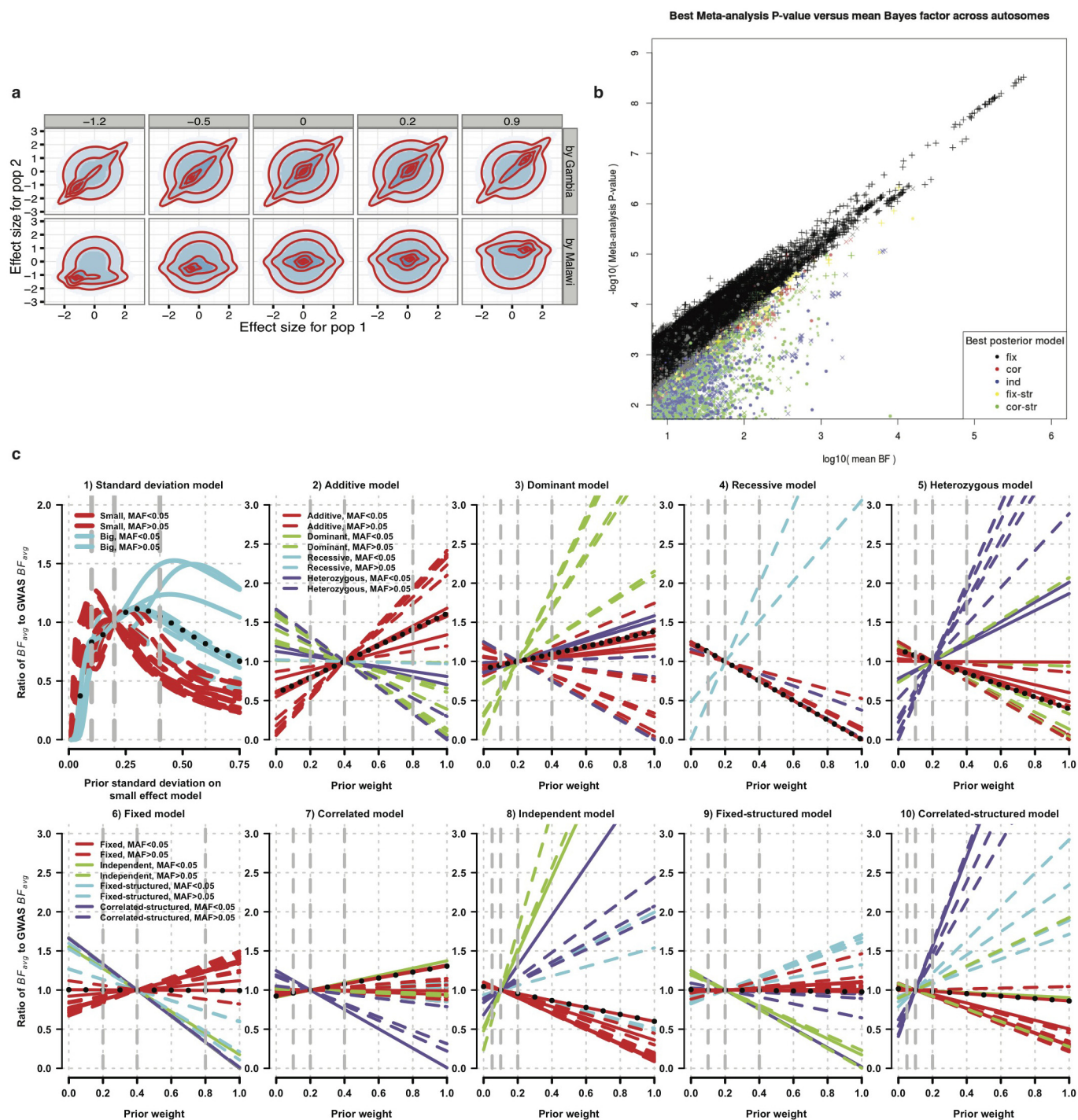
Extended Data Figure 3 | Imputation performance. **a, b**, Empirical distribution of concordance and accuracy (r^2) between typed and re-imputed SNPs in the three discovery cohorts. Solid lines represent SNPs with frequency below 5% and dashed lines represent SNPs with frequency of at least 5%. **c**, Per-sample concordance and accuracy (type 0 r^2) across the whole

genome, as estimated by re-imputing genotyped SNPs. Values are averaged over imputation chunks. **d**, Average accuracy between genotype and re-imputed SNPs in each cohort, plotted against frequency, in 1% frequency bins.



Extended Data Figure 4 | Detail of population structure. **a–c**, Top ten principal components (PCs) in Gambia (**a**), Malawi (**b**) and Kenya (**c**). Where ethnicity was reported, points are coloured by ethnicity for ethnicities with at least 50 samples. **d**, Logistic regression *P* values and direction of effect for the top ten PCs on severe malaria status in each cohort. **e**, *q–q* plots for additive model association test *P* values in Gambia, Malawi, Kenya, and for fixed-effect meta-analysis. Dashed lines represent the 99% CI computed

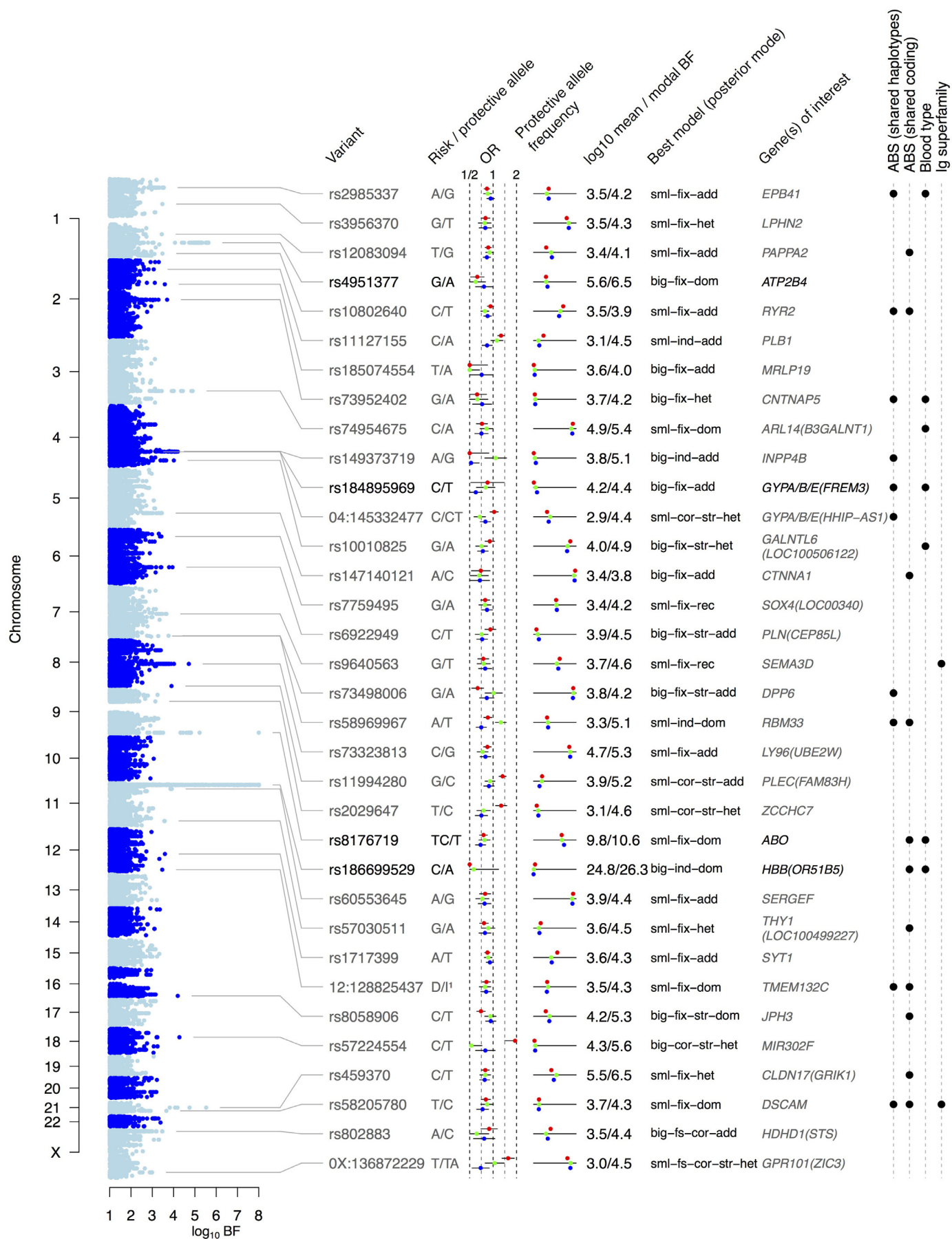
marginally at each variant. Circles and points represent points lying respectively outside and inside the 99% CI. **f**, Comparison of association test *P* values for logistic regression (SNPTEST, *x*-axis) and linear mixed model (MMM, *y*-axis) for Gambia, Malawi, Kenya, and for fixed-effect meta-analysis. Variants in tier 1 are coloured blue, with the lead marker at the *FREM3/GYPE* region coloured red.



Extended Data Figure 5 | Detail of Bayesian analysis of discovery cohorts.

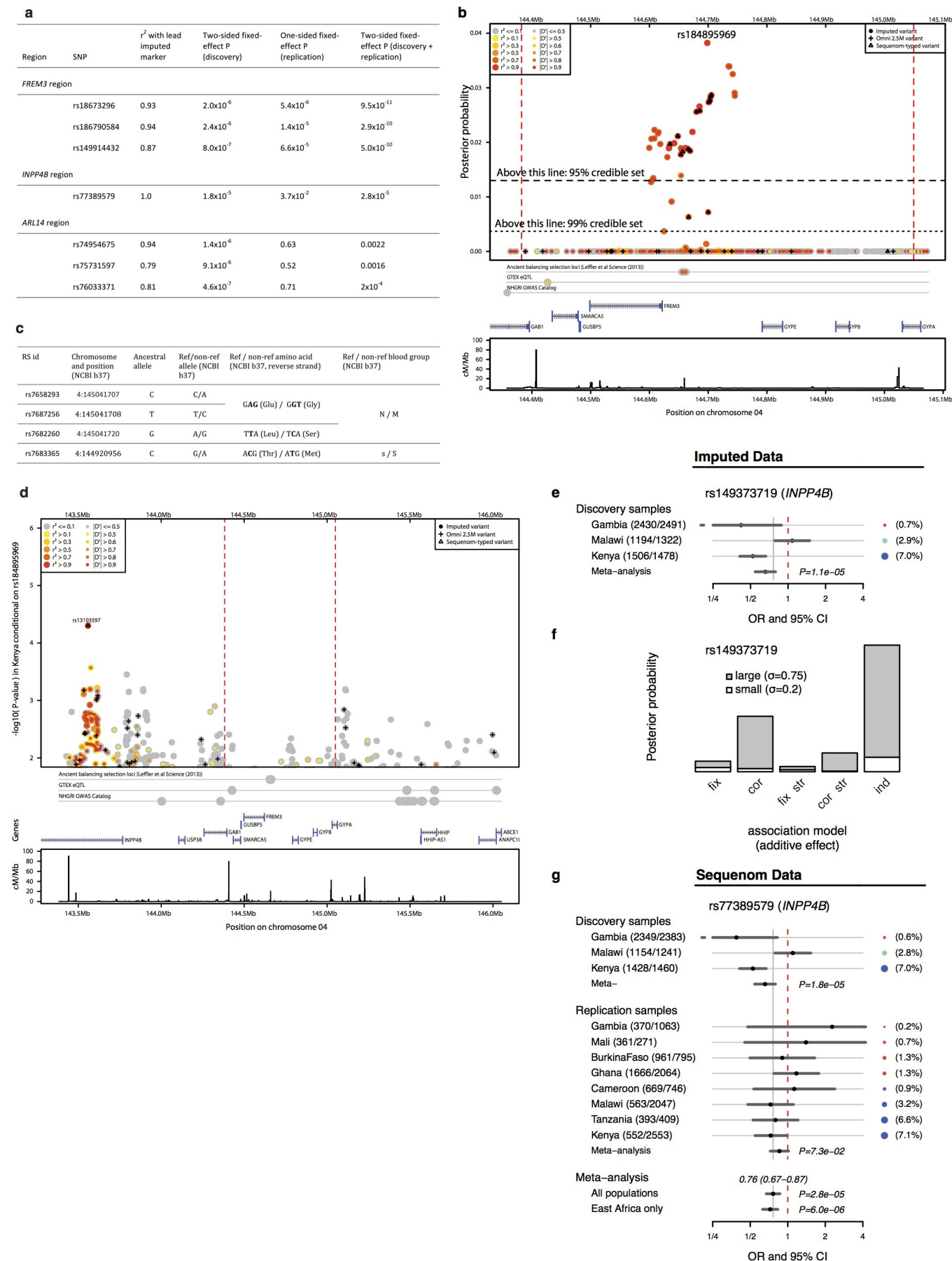
a, Visualization of slices through the combined prior on effect sizes in three cohorts for mode-of-inheritance-specific models. Top row: slices through the prior effect size on Kenya (x-axis) and Malawi (y-axis) for constant effect size in The Gambia (panels). Bottom row: slices through the prior effect size on Kenya (x-axis) and The Gambia (y-axis) for constant effect size in Malawi (panels). Red lines represent a factor of 10 in the prior density. **b**, Comparison of BF_{avg} (x-axis) with the minimum fixed-effect meta-analysis P value minimized across additive, dominant, recessive or heterozygote modes of inheritance (y-axis). Values are plotted on \log_{10} and $-\log_{10}$ scales. Colour indicates the heterogeneity model of the model with the highest posterior weight. **c**, Sensitivity of BF_{avg} to changes in prior. Plots show BF_{avg} ratio (y-axis) plotted against one-dimensional parameterisations of the prior (x-axis), for the 32 autosomal SNPs in tier 1. Solid lines represent variants with $MAF < 5\%$

averaged across populations, and dashed lines represent variants with $\text{MAF} \geq 5\%$. Black dots indicate the lead marker at the *FREM3/GYPE* locus. Colour indicates the effect size, mode of inheritance, or heterogeneity model for the model with highest posterior weight under the GWAS prior. Dashed grey vertical lines indicate the x-axis value corresponding to the prior used in the GWAS, and one-half and twice that value. Plots are parameterized by (1) the prior standard deviation of the small-effect model keeping the prior standard deviation of the large and small-effect models in the ratio 0.75:0.2; (2–5) the prior weight on additive, dominant, recessive or heterozygote modes of inheritance; (6–10) the prior weight on fixed, correlated, independent, fixed-structured and correlated-structured models. For each parameterization prior, weights on other models are kept in the same relative proportion. For further details see Supplementary Note 4.



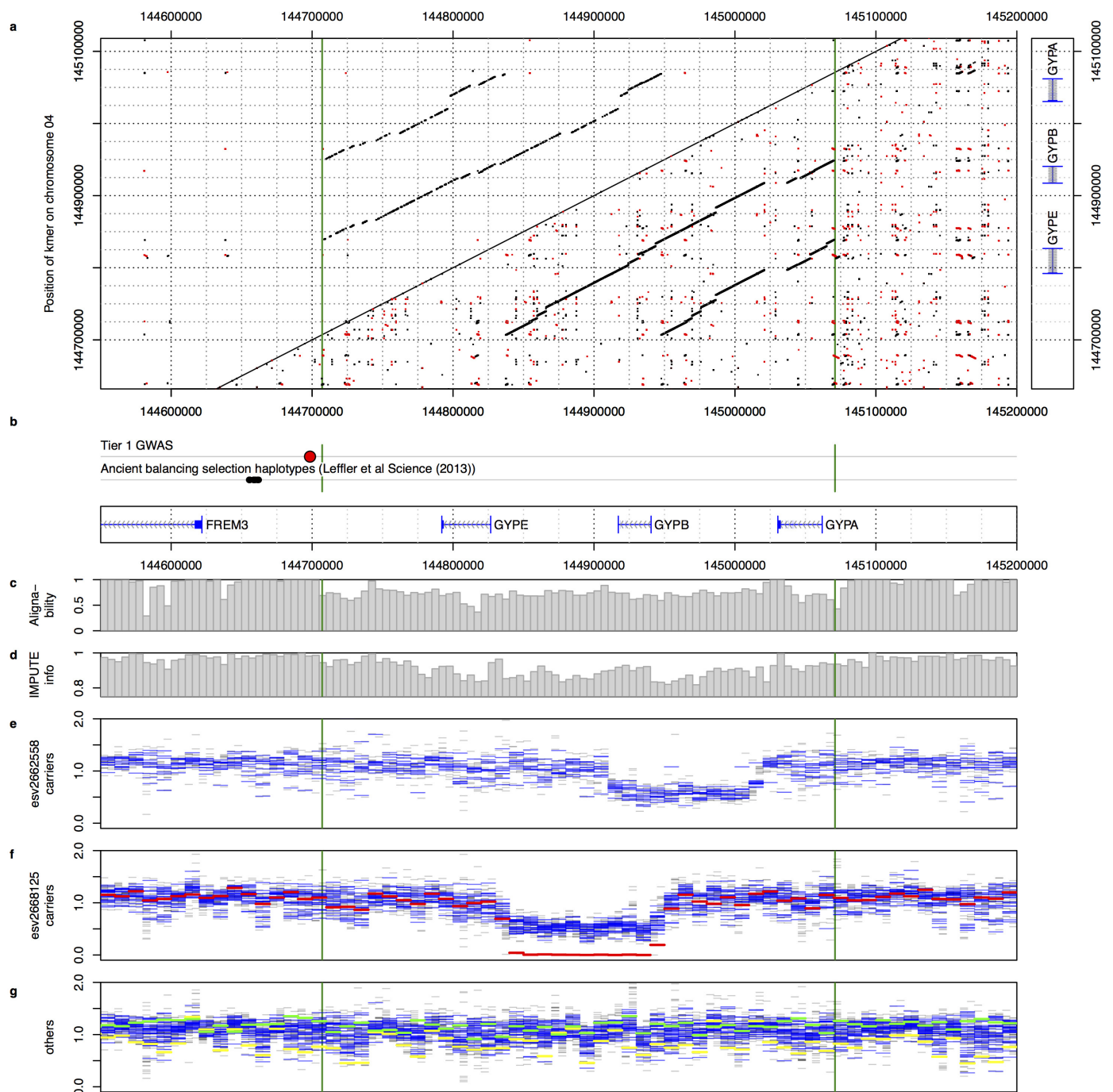
Extended Data Figure 6 | Strongest regions of association in the Bayesian analysis of the three discovery cohorts. Plot on left shows the \log_{10} model-averaged Bayes Factor (BF_{avg}). Table shows the SNP with the highest BF_{avg} in each region (lead SNP), gene(s) of interest in the region, the model with the highest posterior weight at the lead SNP and its BF. Models are described in Methods; sml, big: small- and large-effects models; fix, cor, ind, fix-str, cor-str: fixed, correlated, independent, fixed-structured and correlated-structured

effect models; add, dom, rec, het: additive, dominant, recessive, heterozygote models; fs: fixed-between-sex model. Coloured points indicate the OR and the protective allele frequency in Gambia (red), Malawi (green) and Kenya (blue). The right-hand columns indicate regions containing shared chimpanzee–human haplotypes or coding SNPs⁴ (ABPs), blood group genes, or immunoglobulin superfamily genes.



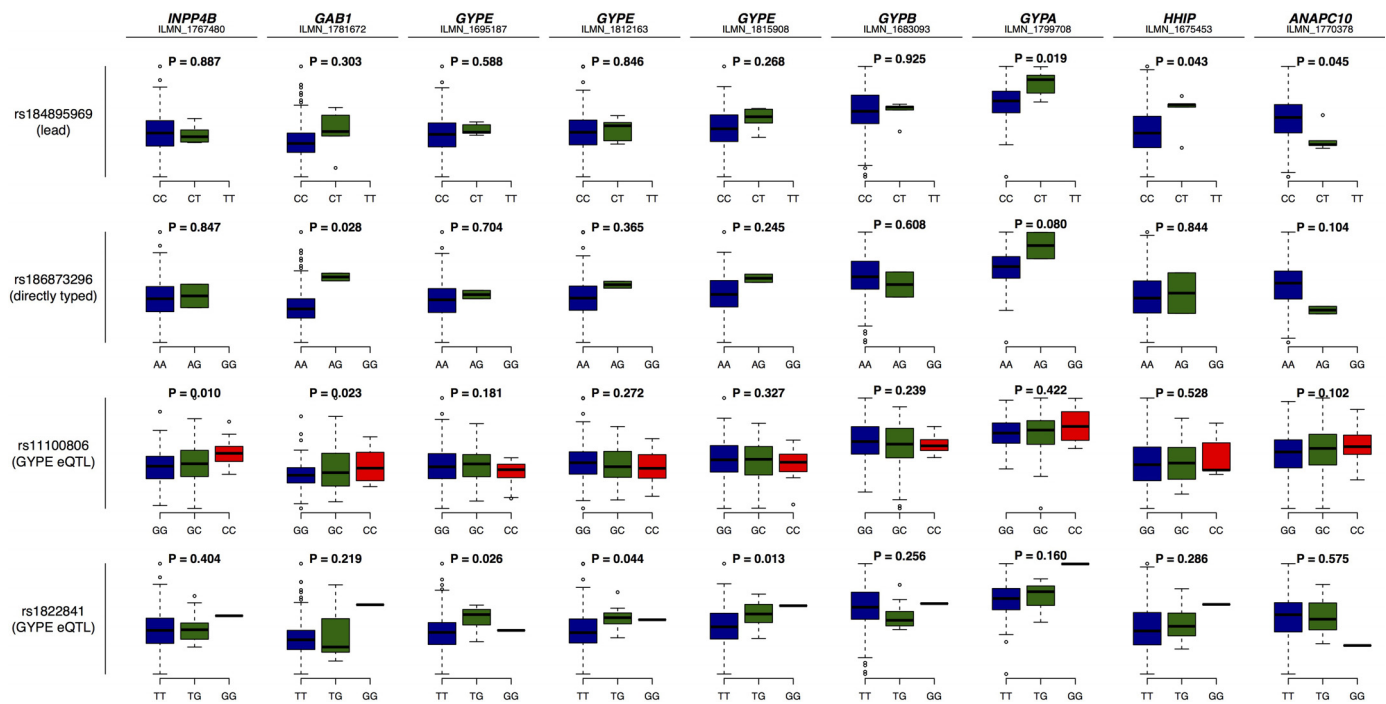
Extended Data Figure 7 | Evidence for association in the *FREM3/GYPE*, *INPP4B* and *ARL14* regions. **a**, Evidence for association at directly typed SNPs. **b**, Posterior probability that variants in the *FREM3/GYPE* region are causal, assuming a single variant in the region is causal⁵⁰, based on the BF_{avg} for typed and imputed variants. Dashed lines indicate the 95% and 99% credible sets. See Fig. 1 legend for further details. **c**, Details of SNPs encoding the common MNS blood groups. Coordinates and alleles are with respect to the NCBI b37 human reference sequence. **d**, Evidence for possible independence of effects at the *FREM3* and *INPP4B* loci in Kenya by conditional analysis. *y*-Axis

represents $-\log_{10}(\text{association } P \text{ value})$ conditional on the imputed dosage at rs184895969. Points are coloured by linkage disequilibrium with the top SNP rs13103597. **e**, Forest plot showing sample size, estimated OR and 95% CI for the lead imputed SNP (rs149373719) in *INPP4B* under an additive model of association. **f**, Bar plot showing the posterior weight on different models of heterogeneity at rs149373719 under the prior used in the GWAS, assuming an additive model of association. **g**, Forest plot showing evidence in both discovery and replication samples in the Sequenom data at rs77389579 in *INPP4B*. See Fig. 2 legend for further details.



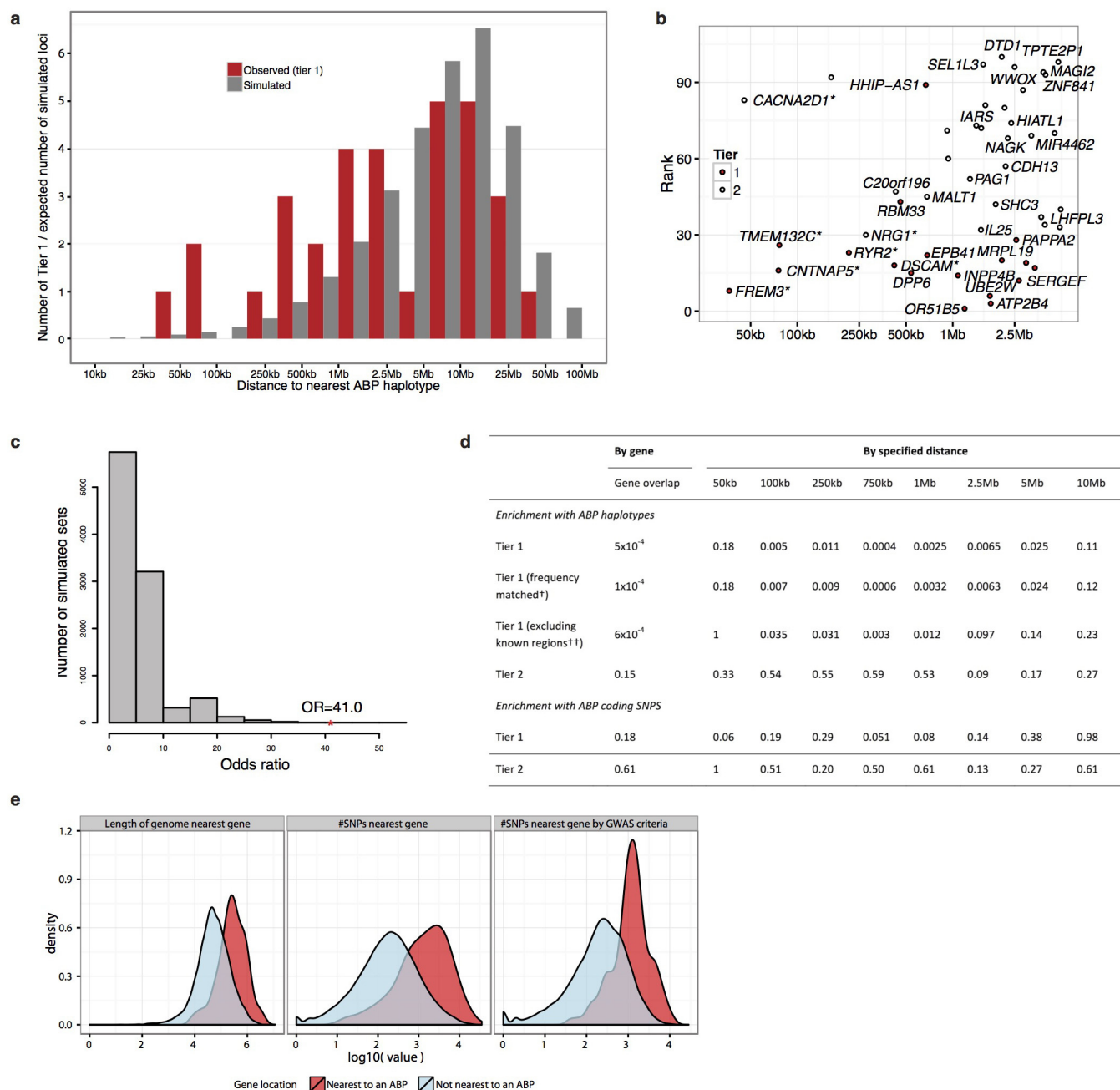
Extended Data Figure 8 | Sequence homology, alignability and structural variation in the glycophorin region. **a**, Co-occurrence of short segments of DNA (k -mers; top triangle: $k = 100$, bottom triangle: $k = 25$) in the human reference sequence. Each point represents a k -mer that maps to the locations indicated by the x - and y -axis positions, either on the same strand (black points) or opposite strands (red points). Green vertical lines in this and subsequent panels delineate the region of high homology surrounding the three glycophorins. **b**, The location of the lead GWAS marker, ABPs, and protein-coding genes in the region. **c**, Alignability of the 100-mer at each position of the reference, up to two mismatches. Values are taken from the UCSC genome browser mappability track and averaged over 5 kb bins. **d**, IMPUTE info measure in Kenya for variants with frequency at least 5%, averaged over 5 kb bins. **e–g**, Coverage for samples from YRI and LWK in 1000 Genomes Project Phase I carrying *esv2662558*, carrying *esv2668125*, or not carrying either

deletion, respectively. Coverage for each individual is normalized by the mean coverage for that individual across chromosome 1, and only computed at positions with alignability = 1 for all 100-mers overlapping the position, and for reads with mapping quality at least 20. Values are averaged over 5 kb (grey) and 10 kb (blue) bins. Three samples with apparently erroneous calls in the 1000 Genomes Phase I genotype release are coloured (NA18519, red; NA19185, yellow; NA19222, green) and assigned to the status indicated by their coverage profile. Panel **g** represents a random sample of 30 individuals not carrying the deletion in addition to the two with erroneous genotype calls. Coverage computation was performed using the BAM files available from the 1000 Genomes Project in October 2014, downloaded from <ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data>. Four African samples in the Phase I release were not assessed because they are not included in this directory.



Extended Data Figure 9 | Correlation between the genotypes at SNPs of interest within the *GYPE/A/B* locus and reported gene transcription levels in samples from the YRI and LWK HapMap cohorts¹⁷. P values are for a

trend test of association where more than one genotype class is present. Only assays targeting the glycoporphins, and those with a P value below 0.05 are shown.



Extended Data Figure 10 | Detail of enrichment analysis. **a**, Red histogram: the empirical distribution of the \log_{10} distance of observed tier 1 loci to the nearest ABP haplotype. Grey histogram, distribution of distances for 10,000 simulated tier 1 sets. **b**, The \log_{10} distance of tier 1 (filled red circles) and tier 2 (empty circles) loci to the nearest ABP, plotted against their rank in BFav order (stronger signals have lower rank). Loci are annotated with the nearest gene where a gene exists within the association region. Asterisks denote nearest genes that are also the nearest gene to an ABP shared haplotype. **c**, Empirical null distribution of the odds ratio for the enrichment of tier 1 loci in the set of genes closest to an ABP shared haplotype, based on 10,000 simulated SNP sets. The red asterisk and text indicate the odds ratio for the observed tier 1 loci. **d**, Distribution of the proportion of the genome which identifies a given gene as nearest, for genes in or not in the set annotated as

nearest an ABP haplotype. Left, distribution of the length of the genome for which the given gene is unambiguously the closest gene. Middle, distribution of the number of SNPs in our study for which the given gene is the closest gene. Right, distribution of the number of SNPs in our study for which the given gene is the nearest gene within a recombination interval of $2.5 \text{ cM} \pm 25 \text{ kb}$ around the SNP, as used to determine nearest genes to GWAS lead SNPs. **e**, Empirical P values for enrichment of ABP haplotypes and coding SNPs in tier 1 and tier 2 GWAS regions. Second column, P values for enrichment by gene overlap. Third to tenth column, P values for enrichment by proximity at different length scales. †Results for simulations using SNPs frequency-matched to GWAS tier 1 loci in 1% frequency bins. ††Results for simulations excluding the regions of *ABO*, *HBB*, *ATP2B4*, *FREM3*, *INPP4B*, and *HHIP-AS1*.

Plasticity-driven individualization of olfactory coding in mushroom body output neurons

Toshihide Hige^{1†}, Yoshinori Aso², Gerald M. Rubin² & Glenn C. Turner¹

Although all sensory circuits ascend to higher brain areas where stimuli are represented in sparse, stimulus-specific activity patterns, relatively little is known about sensory coding on the descending side of neural circuits, as a network converges. In insects, mushroom bodies have been an important model system for studying sparse coding in the olfactory system^{1–3}, where this format is important for accurate memory formation^{4–6}. In *Drosophila*, it has recently been shown that the 2,000 Kenyon cells of the mushroom body converge onto a population of only 34 mushroom body output neurons (MBONs), which fall into 21 anatomically distinct cell types^{7,8}. Here we provide the first, to our knowledge, comprehensive view of olfactory representations at the fourth layer of the circuit, where we find a clear transition in the principles of sensory coding. We show that MBON tuning curves are highly correlated with one another. This is in sharp contrast to the process of progressive decorrelation of tuning in the earlier layers of the circuit^{2,9}. Instead, at the population level, odour representations are reformatted so that positive and negative correlations arise between representations of different odours. At the single-cell level, we show that uniquely identifiable MBONs display profoundly different tuning across different animals, but that tuning of the same neuron across the two hemispheres of an individual fly was nearly identical. Thus, individualized coordination of tuning arises at this level of the olfactory circuit. Furthermore, we find that this individualization is an active process that requires a learning-related gene, *rutabaga*. Ultimately, neural circuits have to flexibly map highly stimulus-specific information in sparse layers onto a limited number of different motor outputs. The reformatting of sensory representations we observe here may mark the beginning of this sensory-motor transition in the olfactory system.

The Kenyon cell (KC) axons are arranged in parallel bundles that form the output lobes of the mushroom body (MB). Mushroom body output neurons (MBONs) extend dendrites into those lobes, with different MBON types innervating distinct subregions^{7,8} (Fig. 1a). We expressed the calcium sensor GCaMP5 in MBONs using a series of split-GAL4 lines⁸ (Extended Data Table 1) and measured odour tuning using *in vivo* two-photon imaging, quantifying response magnitude as the area under the $\Delta F/F$ curves (F , fluorescence intensity; Fig. 1b). The high specificity of these drivers typically enabled us to track activity of individual MBONs. We thereby successfully collected data from 17 types/combination of types of MBONs, covering 18 out of 21 cell types, and one additional MBON from the calyx (Fig. 1c and Extended Data Figs 1 and 2; see Methods).

Consistent with high convergence at this stage of the circuit^{7,8}, MBONs were generally broadly tuned to odours, as observed in other insects^{10–12}, although there were a few exceptions (for example, $\alpha 2p3p$, $\beta'1$ and MB-CP1 neurons; Extended Data Fig. 3). In the MBONs with axonal projections inside the MB lobes ($\beta 1$, $\gamma 1pedc$, and $\gamma 4$ neurons), we observed prolonged rise times (Extended Data Fig. 4).

One of the important factors governing the stimulus-specificity of population-level representations is how independent and decorrelated

their sensory tuning is. Optimal coding theory dictates that a compact neuronal population most efficiently conveys stimulus-specific information if the tuning properties of different neurons are decorrelated, so the redundancy of their signalling is minimized¹³, which we refer to as tuning decorrelation. We confine our analysis here to a tuning-curve-based view of the system, and do not explore the role that temporal patterning of spikes might have in conveying information, as has been shown in other systems^{11,14}. Overall, odour tuning of the MBON population was notable for its lack of diversity, showing high levels of correlation (Figs 1d and 2e). We found no obvious relationship between the degree of tuning correlation of different MBONs and their type of input KC, the neurotransmitter they release, or where they subsequently project (Fig. 1d and Extended Data Fig. 5a, b).

These highly correlated, dense response patterns were in sharp contrast to the KCs. The calcium responses of KCs to the same set of odours, measured at the cell body layer, were sparse and specific (Fig. 2a, b), with much lower levels of tuning correlation (Fig. 2e). To visualize how odour representations are transformed between the KCs and MBONs, we used principal component analysis (PCA) to represent population response patterns observed on each stimulus trial (Fig. 2c; see Methods and Extended Data Fig. 6). Although different odour clusters were well-separated in the KCs, in MBONs they were much closer to one another and often partially overlapping. Nevertheless, there was a coarse structure to the distribution of different odours, and some were well-separated. This basic structure was conserved when we analysed subpopulations of MBONs according to their axonal projection sites (Extended Data Fig. 5c). The close proximity of odour clusters visualized by PCA was reflected in a lower score of odour classification analysis in MBONs than KCs (Fig. 2d; see Methods). Importantly, this was not simply caused by the sharp reduction in the number of neurons, or their broad tuning compared to the KCs. When we held cell number and tuning breadth constant, but artificially decorrelated MBON tuning by assigning rearranged odour labels to each cell's tuning curve, classification accuracy markedly increased (Fig. 2d, e; see Methods). Furthermore, when we examined the number of distinct odour clusters in MBON space, relatively few clusters were apparent, but artificial decorrelation of MBON tuning increased the number of clusters to match the number of odours, just like the KC representations (Fig. 2f; see Methods). These results clearly show that a neuronal population of this size and breadth of tuning is capable of representing odour identity accurately; however, the correlations in MBON tuning properties place an important limit on that capacity. We note that it is still possible that specific information about odour identity could be carried in the precise timing of MBON spike trains^{11,14}.

We then asked what features of sensory information become available at this layer. To address this, we calculated the correlations between neural representations of all pairs of odours in KCs and MBONs and compared the distributions (Fig. 2g). In KCs, this distribution showed a single sharp peak near zero, indicating that odour representations are largely decorrelated; in fact, artificially

¹Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA. ²Janelia Research Campus, Howard Hughes Medical Institute, 19700 Helix Drive, Ashburn, Virginia 20147, USA. [†]Present address: Janelia Research Campus, 19700 Helix Drive, Ashburn, Virginia 20147, USA.

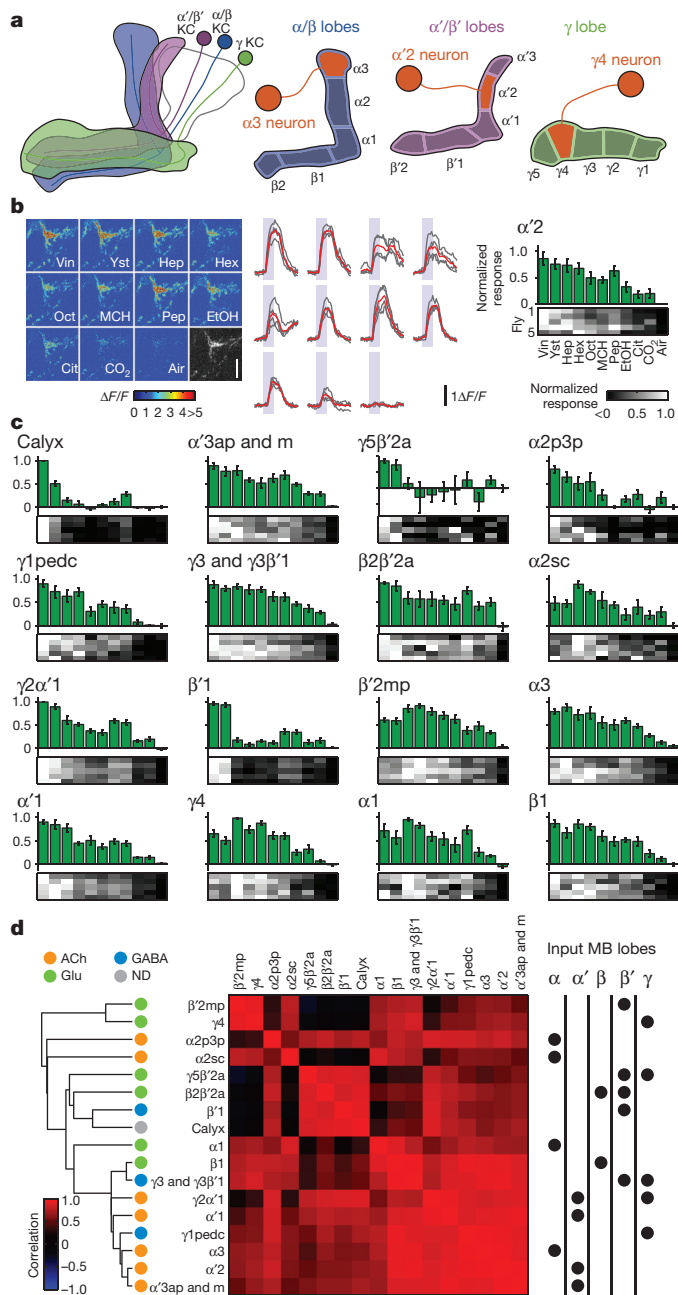


Figure 1 | Summary of olfactory tuning patterns in MBONs. **a**, Schematic of the KCs and MBONs. **b**, GCaMP5 imaging in the $\alpha'2$ neuron. Left, example images of the odour responses from axonal arbor. Greyscale image, baseline fluorescence. Scale bar, 10 μm . Vin, apple cider vinegar; Yst, yeast; Hep, 2-heptanone; Hex, 1-hexanol; Oct, 3-octanol; MCH, 4-methylcyclohexanol; Pep, peppermint; EtOH, ethanol; Cit, citronella. Middle, $\Delta F/F$ time courses in the same cell. Grey, individual trials ($n = 4$). Red, mean. Shaded areas, 2-s odour stimulations. Right, odour tuning profiles. Bars, mean (\pm s.e.m.) normalized tuning ($n = 5$ flies; see Methods). Heat map, tuning in each fly. **c**, Odour tuning in the remaining 16 MBON cell types. **d**, Correlation matrix of tuning patterns of MBONs (Pearson's r ; middle). Cell types are arranged according to the dendrogram (left) based on the correlation distances ($1 - r$). Neurotransmitters (ACh, acetylcholine; Glu, glutamate; GABA, γ -aminobutyric acid; ND, not determined) and the input lobes are indicated by dots.

decorrelating KC tuning had little further effect. By contrast, in MBONs correlation coefficients ranged widely without an obvious peak around zero. Both the dense format of MBON representations, as well as the pattern of activity contribute to this distribution shape,

since artificially decorrelating MBON tuning properties resulted in a wider distribution than in the KCs, but now with a clear peak around zero (Fig. 2g and Extended Data Fig. 5e). Some of the relationships between odours are partially inherited from previous layers of neurons, because we observed a significant positive correlation between correlation coefficients in KCs and in MBONs (Fig. 2i). Notably, negative correlations between different odour representations, a rare feature in KCs, become relatively common in MBONs, especially in MBON subpopulations with axonal projections to particular downstream areas (Fig. 2h and Extended Data Fig. 5d). Thus, it seems that MBONs convey a sense of interrelationship between odours, be it positive or negative. In this context, it is interesting to note that two odour groups of opposite valence were reliably the most distantly located in MBON coding space (Fig. 2c and Extended Data Fig. 5c) and were never misclassified with each other (Extended Data Fig. 7). One of those groups was apple cider vinegar and yeast, which are attractive food odours for *Drosophila*^{15,16}. The other was CO₂ together with citronella, which are both reported to be natural repellents to *Drosophila*^{17,18}.

We next focused on odour tuning properties at the single-cell level. Specifically, we asked whether cell types with uniquely identifiable anatomy also have consistent tuning properties in different animals. Such functional stereotypy is readily apparent in the projection neurons at the second layer of the circuit^{19,20}. In MBONs, by contrast, we observed a range of similarity in tuning across animals depending on cell types (Fig. 3a). Some MBONs showed highly consistent responses; however, several had very diverse tuning. Among all the MBONs, the $\alpha 2sc$ neuron exhibited the greatest variability. This was not due to ambiguity in cell identification, since there is only one $\alpha 2sc$ neuron per hemisphere⁸. Moreover, a similar level of inter-animal variability was also observed with whole-cell recordings (Fig. 3b–d).

Tuning patterns of individual KCs are not stereotyped across animals²⁰, which is likely to arise from the probabilistic input connectivity of the projection neurons^{21,22}. Is the variability in MBON tuning simply inherited from the probabilistic organization of the previous layer? If so, tuning properties of MBONs should be as variable across hemispheres in the same brain as they are across different animals. However, this was clearly not the case. Pairs of $\alpha 2sc$ neurons from opposite sides of the same brain exhibited strikingly similar tuning compared to those from different brains (Fig. 3e and g). We observed similar results in three other MBON types (Extended Data Fig. 8). Thus, tuning patterns of MBONs are individual-specific as a result of a process that is coordinated across hemispheres, rather than random wiring patterns.

We next set out to ask how this functional individuality of the circuit arises at the level of MBONs. The dense dendritic arbor of the MBONs implies that MBONs summate input from many KCs^{7,23}. If so, does the variable tuning of MBONs derive from variable overall levels of population activity in KCs? However, while tuning profiles of individual KCs are not predictable^{20–22}, the summed output from the overall KC population could be consistent across animals, since projection neurons from different glomeruli have relatively stereotyped numbers of output terminals in the MB, suggesting that the total excitatory drive to the KCs may be characteristic for each odour^{24,25}. Furthermore, the number of responding KCs for a given odour is positively correlated with the total activity of olfactory receptor neurons responding to that odour³. To directly examine the variability of bulk MB output, we imaged calcium responses in the KC axon bundle at the site where it contacts the $\alpha 2sc$ neuron. We found that the tuning of the bulk KC population was relatively consistent from fly to fly, in contrast to the $\alpha 2sc$ neuron (Fig. 3f, g). We thus found no sign of individuality in the summed activity of the KCs. But then, why do MBONs, thought to receive heavily convergent input from KCs, not end up with stereotyped tuning patterns? To better understand how KC activity is integrated by MBONs, we measured functional connectivity between α/β KCs and $\alpha 2sc$ neurons by paired whole-cell recordings (Fig. 4a, b). Surprisingly, we found only 7 pairs out of 24 with an excitatory

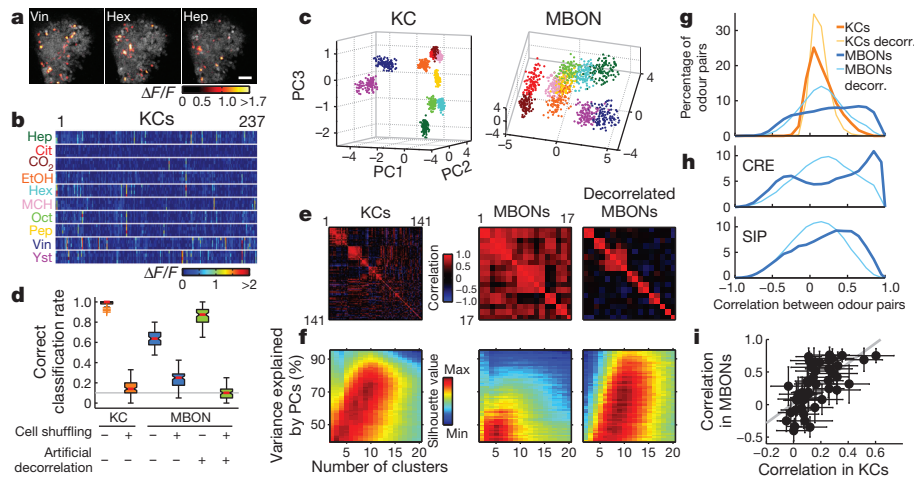


Figure 2 | Transformation of population representations from KCs to MBONs. **a**, Representative odour responses of KC somata in a single fly with pseudo-coloured $\Delta F/F$ images overlaid on greyscale basal fluorescence. Scale bar, 20 μm . **b**, Summary of responses from the fly in **a**. Columns, KCs. Rows, trials sorted by odours. **c**, Odour representations in a single representative fly (KCs, same fly as **a**) or virtual fly (MBONs) projected onto the first three principal component axes (PC; 100 data points per odour; see Methods). Colours indicate odours as in **b**. **d**, Accuracy of odour classification (100 classification scores for both KCs and MBONs; see Methods). Red lines indicate medians, boxes interquartile ranges, notches 95% confidence intervals, whiskers data ranges, and crosses outliers. Grey line, chance level. **e**, Correlation matrices of tuning curves. Data are from a single representative fly (KCs) or virtual fly (MBONs). **f**, Clustering analysis of population activity patterns in

KCs and MBONs (KCs, $n = 10$ flies; MBONs, $n = 20$ virtual flies; see Methods). Highest silhouette values indicate optimal clustering. Silhouette value ranges are KCs, 0.37–0.84; MBONs, 0.35–0.74; decorrelated MBONs, 0.27–0.65. **g**, Correlation coefficients (Pearson's r) of neural representations of ten odours (45 odour pairs), calculated using the response patterns averaged across trials in KCs ($n = 10$ flies, orange) and MBONs ($n = 1,000$ virtual flies, blue). Mean distributions are shown. Distributions of artificially decorrelated data are in lighter colours. **h**, Same as **g**, but for MBON subpopulations with axonal projections to different areas. See also Extended Data Fig. 5. CRE, crepine; SIP, superior intermediate protocerebrum. **i**, Correlation coefficients in KCs and MBONs are positively correlated. Each dot corresponds to correlation coefficient of a specific odour pair (mean \pm s.d.). Grey line, linear regression ($R^2 = 0.42$, $P < 10^{-5}$).

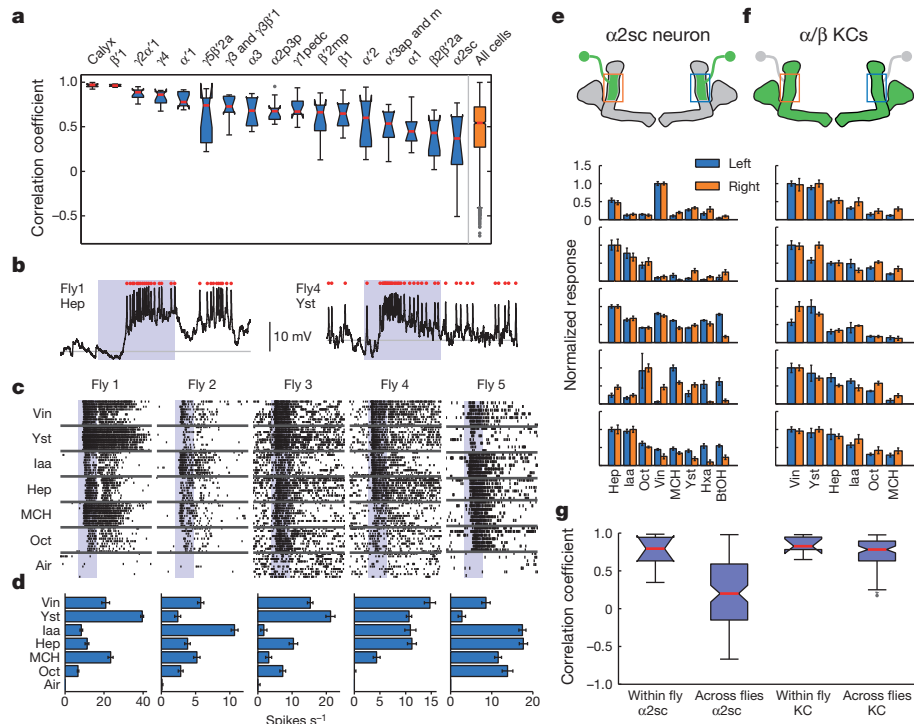


Figure 3 | Individualization of tuning in MBONs. **a**, Correlations of MBON tuning patterns across different flies ($n = 5$ flies for each cell type). **b**, Sample traces from whole-cell recordings from $\alpha 2\text{sc}$ neurons. Red dots, spikes. Shaded areas, 1-s odour presentations. **c**, Raster plots of $\alpha 2\text{sc}$ odour responses. Iaa, isoamyl acetate. **d**, Tuning curves (mean \pm s.e.m.) from **c**. **e**, Odour tuning of pairs of $\alpha 2\text{sc}$ neurons on the left and right hemispheres of five representative flies, from GCaMP5 imaging (normalized to the strongest response, mean \pm s.e.m.). BtOH, 1-butanol; Hxa, 1-hexanol. **f**, α/β KC population tuning

across different hemispheres of the same fly, recorded at the $\alpha 2$ region of the MB lobes. **g**, Tuning patterns of $\alpha 2\text{sc}$ neurons from different hemispheres of the same fly are more correlated than those from different flies ($n = 8$ flies, $P < 10^{-4}$, Tukey's post-hoc test following two-way analysis of variance (ANOVA)). Correlations of KC population tuning were indistinguishable within versus across flies ($n = 7$ flies, $P > 0.95$, Tukey's post-hoc test). Interactions between cell types ($\alpha 2\text{sc}$ versus KCs) and comparison types (within versus across flies) were significant ($P < 0.05$, two-way ANOVA).

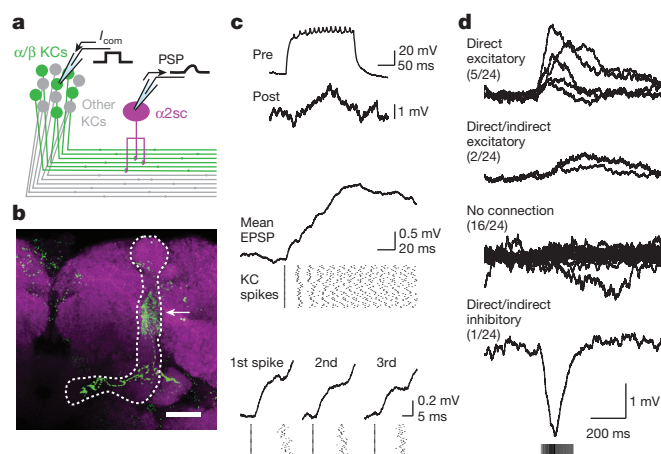


Figure 4 | Connection probability between KCs and a MBON. **a**, Schematic of paired whole-cell recording. **b**, Post-hoc immunohistochemistry of biocytin infused via whole-cell electrodes (green, maximum intensity projection). Axon of α/β KC, extending within α/β lobes (dotted line), and dendritic arbor of $\alpha 2sc$ neuron (arrow) are visible. Magenta, anti-nc82 staining. Scale bar, 20 μm . **c**, A directly connected KC- $\alpha 2sc$ neuron pair. Top, a train of KC spikes (Pre) and the evoked response in a $\alpha 2sc$ neuron (Post) recorded simultaneously. Middle, discrete excitatory postsynaptic potential (EPSP) steps visualized by spike-trigger-averaging MBON membrane potential on the first KC spike of each trial. Bottom, enlarged spike-trigger-averaged EPSPs for the first, second and third spikes in the train reveal a short synaptic delay (< 2 ms). **d**, The four categories of synaptic connectivity (see Methods). The first-spike-trigger-averaged responses for all pairs in each category are overlaid. Timing of the KC spikes is indicated by greyscale bar, whose intensity shows the mean relative spike rate.

connection, only 5 of which were likely to be monosynaptic (Fig. 4c, d and Extended Data Fig. 9; see Methods). This gives a probability of connection of $< 30\%$, far more selective than the all-to-one convergence suggested by the dendritic anatomy. Thus, our results suggest that $\alpha 2sc$ neurons are capable of extracting very different information in different animals, even from presynaptic KCs that have similar overall population tuning, through individual-specific connectivity with KCs.

One process that could plausibly underlie such flexible wiring is synaptic plasticity; indeed plasticity has been observed at this synapse in other insects²⁶. To test this we examined whether *Rutabaga*, an adenylyl cyclase required for learning⁴, is involved in generating the cross-fly differences in tuning. In *rutabaga* mutants, across-fly tuning variability of the $\alpha 2sc$ neurons was markedly reduced, and there was no longer a significant difference between within- and across-fly correlations (Fig. 5). Thus the cross-fly differences in this neuron are the result of an active process that requires *rutabaga* signalling. However, *rutabaga*-dependent plasticity does not seem to be the sole determinant of the MBON tuning because the relatively stereotyped MBON tuning in the mutants is still different from the bulk KC tuning (Fig. 3f). *rutabaga* may also contribute to the coordination of tuning across hemispheres, since within-animal correlations in the mutants tend to be lower than in controls, although this difference was not statistically significant.

This work presents the most complete population-level characterization of tuning at any layer of the olfactory system. This extensive coverage of the population, combined with our back-to-back comparison to the previous layer, enabled us to demonstrate that the progressive decorrelation of neuronal tuning that marks the ascending layers of the sensory circuit^{2,9} comes to an end immediately downstream of the KCs, as the network converges. Fine temporal patterns of spiking, or subtle correlations in signalling across multiple MBONs within each animal, neither of which we could detect here, could contribute to the specificity of odour representations at this layer (but see Extended Data Fig. 6). Nevertheless, our results clearly show that positive and negative correlations arise in MBONs, clustering some odour representations

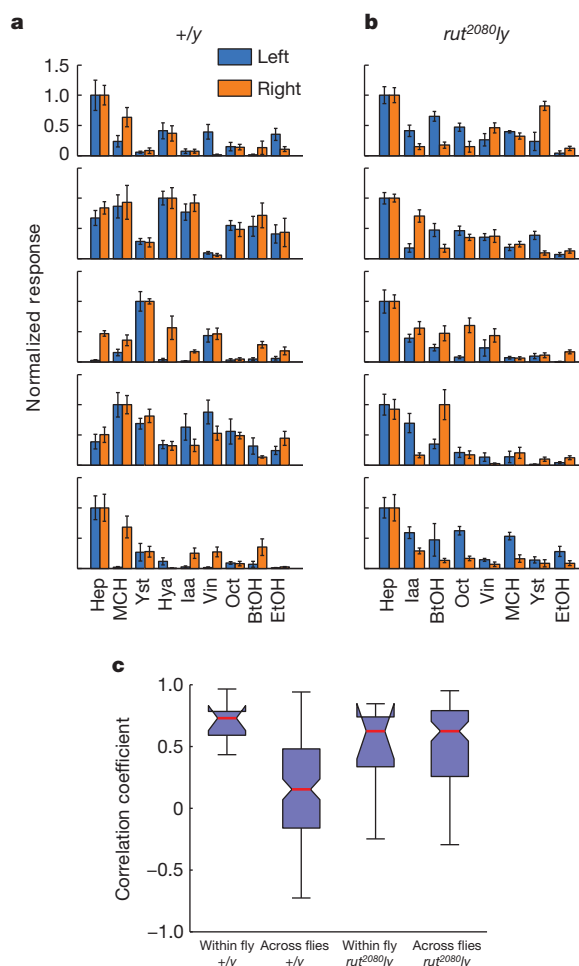


Figure 5 | Cross-individual variability is lost in *rutabaga* mutants. **a**, Odour tuning of a pair of $\alpha 2sc$ neurons on the left and right hemispheres in the same fly, recorded with GCaMP5 imaging (mean \pm s.e.m.). Data from five wild-type males. **b**, Same as **a** but *rut²⁰⁸⁰* males. **c**, Control males show higher variability across flies than within flies ($n = 9$ flies, $P < 10^{-4}$, Tukey's post-hoc test following two-way ANOVA). This difference is lost in *rut²⁰⁸⁰* hemizygous males, which show similar levels of variability both within and across flies ($n = 8$ flies, $P > 0.995$, Tukey's post-hoc test). Interactions between the genotypes and comparison types (within versus across flies) were significant ($P < 0.005$, two-way ANOVA). Within-fly tuning correlations were not statistically different between the two genotypes ($P > 0.65$, Tukey's post-hoc test).

together and pushing others apart. This grouping might be useful when it comes to making a behavioural choice, since the general categorization of stimuli, rather than detailed, stimulus-specific information would be more important at this stage. Thus, MBON representations may be well-suited to control motor outputs²⁷. Interestingly, similarly prominent network convergence occurs with cortical projections to the basal ganglia²⁸, whose main function is to select an appropriate action plan by interpreting the available sensory information.

The plasticity-driven individualized tuning of MBONs is a counterpoint to the highly stereotyped responses of the output neurons of the lateral horn (LH), an olfactory centre that lies in parallel with the MB²⁹ and is implicated in innate responses to odour³⁰. In contrast, the influence of plasticity on MBON tuning highlights the flexibility of this branch of the circuit, where odour representations could be reshaped either to fine tune, or perhaps to override the innate responses driven by the LH pathway, reflecting each fly's individually unique olfactory experience.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 16 July 2014; accepted 11 August 2015.

Published online 30 September 2015.

- Perez-Orive, J. *et al.* Oscillations and sparsening of odor representations in the mushroom body. *Science* **297**, 359–365 (2002).
- Turner, G. C., Bazhenov, M. & Laurent, G. Olfactory representations by *Drosophila* mushroom body neurons. *J. Neurophysiol.* **99**, 734–746 (2008).
- Honegger, K. S., Campbell, R. A. A. & Turner, G. C. Cellular-resolution population imaging reveals robust sparse coding in the *Drosophila* mushroom body. *J. Neurosci.* **31**, 11772–11785 (2011).
- Davis, R. L. Olfactory memory formation in *Drosophila*: from molecular to systems neuroscience. *Annu. Rev. Neurosci.* **28**, 275–302 (2005).
- Campbell, R. A. A. *et al.* Imaging a population code for odor identity in the *Drosophila* mushroom body. *J. Neurosci.* **33**, 10568–10581 (2013).
- Lin, A. C., Bygrave, A. M., de Calignon, A., Lee, T. & Miesenböck, G. Sparse, decorrelated odor coding in the mushroom body enhances learned odor discrimination. *Nature Neurosci.* **17**, 559–568 (2014).
- Tanaka, N. K., Tanimoto, H. & Ito, K. Neuronal assemblies of the *Drosophila* mushroom body. *J. Comp. Neurol.* **508**, 711–755 (2008).
- Aso, Y. *et al.* The neuronal architecture of the mushroom body provides a logic for associative learning. *eLife* **3**, e04577 (2014).
- Olsen, S. R., Bhandawat, V. & Wilson, R. I. Divisive normalization in olfactory population codes. *Neuron* **66**, 287–299 (2010).
- Li, Y. & Strausfeld, N. J. Morphology and sensory modality of mushroom body extrinsic neurons in the brain of the cockroach, *Periplaneta americana*. *J. Comp. Neurol.* **387**, 631–650 (1997).
- MacLeod, K., Bäcker, A. & Laurent, G. Who reads temporal information contained across synchronized and oscillatory spike trains? *Nature* **395**, 693–698 (1998).
- Cassenaer, S. & Laurent, G. Conditional modulation of spike-timing-dependent plasticity for olfactory learning. *Nature* **482**, 47–52 (2012).
- Barlow, H. in *Current Problems in Animal Behaviour* (ed. Thorpe, W. H.) 331–360 (Cambridge Univ. Press, 1961).
- Gupta, N. & Stopfer, M. A temporal channel for information in sparse sensory coding. *Curr. Biol.* **24**, 2247–2256 (2014).
- Semmelhack, J. L. & Wang, J. W. Select *Drosophila* glomeruli mediate innate olfactory attraction and aversion. *Nature* **459**, 218–223 (2009).
- Knaden, M., Strutz, A., Ahsan, J., Sachse, S. & Hansson, B. S. Spatial representation of odorant valence in an insect brain. *Cell Reports* **1**, 392–399 (2012).
- Suh, G. S. B. *et al.* A single population of olfactory sensory neurons mediates an innate avoidance behaviour in *Drosophila*. *Nature* **431**, 854–859 (2004).
- Kwon, Y. *et al.* *Drosophila* TRPA1 channel is required to avoid the naturally occurring insect repellent citronellal. *Curr. Biol.* **20**, 1672–1678 (2010).
- Wilson, R. I., Turner, G. C. & Laurent, G. Transformation of olfactory representations in the *Drosophila* antennal lobe. *Science* **303**, 366–370 (2004).
- Murthy, M., Fiete, I. & Laurent, G. Testing odor response stereotypy in the *Drosophila* mushroom body. *Neuron* **59**, 1009–1023 (2008).
- Caron, S. J. C., Ruta, V., Abbott, L. F. & Axel, R. Random convergence of olfactory inputs in the *Drosophila* mushroom body. *Nature* **497**, 113–117 (2013).
- Gruntman, E. & Turner, G. C. Integration of the olfactory code across dendritic claws of single mushroom body neurons. *Nature Neurosci.* **16**, 1821–1829 (2013).
- Séjourné, J. *et al.* Mushroom body efferent neurons responsible for aversive olfactory memory retrieval in *Drosophila*. *Nature Neurosci.* **14**, 903–910 (2011).
- Wong, A. M., Wang, J. W. & Axel, R. Spatial representation of the glomerular map in the *Drosophila* protocerebrum. *Cell* **109**, 229–241 (2002).
- Marin, E. C., Jefferis, G. S. X. E., Komiyama, T., Zhu, H. & Luo, L. Representation of the glomerular olfactory map in the *Drosophila* brain. *Cell* **109**, 243–255 (2002).
- Cassenaer, S. & Laurent, G. Hebbian STDP in mushroom bodies facilitates the synchronous flow of olfactory information in locusts. *Nature* **448**, 709–713 (2007).
- Aso, Y. *et al.* Mushroom body output neurons encode valence and guide memory-based action selection in *Drosophila*. *eLife* **3**, e04580 (2014).
- Houk, J. C. in *Models of Information Processing in the Basal Ganglia* (eds Houk, J. C., Davis, J. L. & Beiser, D. G.) 3–10 (Massachusetts Institute of Technology, 1994).
- Fişek, M. & Wilson, R. I. Stereotyped connectivity and computations in higher-order olfactory neurons. *Nature Neurosci.* **17**, 280–288 (2014).
- Heimbeck, G., Bugnon, V., Gendre, N., Keller, A. & Stocker, R. F. A central neural circuit for experience-independent olfactory and courtship behavior in *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA* **98**, 15336–15341 (2001).

Acknowledgements We would like to thank V. Jayaraman, J. Dubnau and K. Ito for fly strains. We are grateful to H. Kazama, W. Li and J. Dubnau for helpful advice, and to V. Jayaraman, G. Otazu and the members of the Turner laboratory for valuable comments on the manuscript. This work was supported by NIH grant R01 DC010403-01A1 to G.C.T.; T.H. was partially supported by a Postdoctoral Fellowship for Research Abroad from Japan Society for the Promotion of Science and a Postdoctoral Fellowship from the Uehara Memorial Foundation.

Author Contributions T.H. and G.C.T. designed the experiments with help from Y.A. and G.M.R.; T.H. performed all imaging and electrophysiology experiments and data analyses. Y.A. and G.M.R. created fly strains and collected anatomical data for MBONs. T.H. and G.C.T. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to G.C.T. (turner@cshl.edu).

METHODS

Fly stocks. Flies were raised at room temperature on conventional cornmeal-based medium. All experiments were performed on adult females, 2–5 days post-eclosion, unless otherwise noted. Note that we used the same animal-rearing conditions for all flies when we examined the variability of tuning across different flies, and that we always compared flies of same gender. In several cases, we even recorded on the same day from progeny of the same cross, raised in the same food vial. No randomization or blinding methods were used. For calcium imaging, flies bearing UAS-GCaMP6f (ref. 31; KC cell body imaging) or GCaMP5 (ref. 32; MBON and KC axon imaging) were crossed with appropriate GAL4 or split-GAL4 lines, and the resulting F1 flies were used. We used OK107-GAL4 (refs. 33, 34) for KC imaging (both cell body and lobe imaging) and a series of split-GAL4 lines for MBON imaging⁸ (Extended Data Fig. 2; Extended Data Table 1; see also <http://www.janelia.org/split-gal4> for the expression patterns as well as enhancer fragments used to construct these split-GAL4 lines). For experiments in *rutabaga* mutants, we used the fact that *rutabaga* is X-linked. *rut*²⁰⁸⁰ (ref. 35) females were crossed with UAS-GCaMP5; MB080C males, and their male progeny, hemizygous for *rut*²⁰⁸⁰, were used for imaging. For electrophysiological recording from the α 2sc neuron, we crossed UAS-2eGFP with MZ160-GAL4 (ref. 36). For dual whole-cell recording from α/β KCs and the α 2sc neuron, we crossed UAS-2eGFP; MZ160 with *c739*, an α/β -specific driver^{34,37}. UAS-GCaMP flies were provided by V. Jayaraman, *rut*²⁰⁸⁰ and *c739* by J. Dubnau and MZ160 by K. Ito. UAS-2eGFP was obtained from the Bloomington stock centre.

Nomenclature of MBONs. The set of 21 cell types of MBONs from the MB lobes and one from the calyx are defined by their morphology, having dendrites inside the MB and axonal projections elsewhere⁸. For simplicity, in this paper we call each MBON according to its dendritic region in the MB lobes, such as α 1, γ 5 β '2a and so on (Extended Data Table 1).

Odour delivery. The following monomolecular odorants were purchased from Sigma and used as stimuli: 2-heptanone (Hep; CAS# 110-43-0), 1-hexanol (Hex; 111-27-3), 3-octanol (Oct; 589-98-0), 4-methylcyclohexanol (MCH; 589-91-3), ethanol (EtOH; 64-17-5), isoamyl acetate (Iaa; 123-92-2), 1-hexanol (Hxa; 66-25-1), 1-butanol (BtOH; 71-36-3) and hexyl acetate (Hya; 142-92-7). We also used natural essential oils from peppermint (Pep) and citronella (Cit; Aura Cacia) as well as other natural odours, apple cider vinegar (Vin; Richfood) and yeast (Yst; Lessafre). In imaging experiments, odours were presented through a custom-built device as described previously³. 40-ml vials were loaded with 5-ml pure odorants, except for essential oils, which were diluted with mineral oil at 1:100. Yeast was prepared by adding 5 ml distilled water to 1 g dry yeast. Saturated headspace vapours were diluted by two steps of air dilutions down to 10% (experiments in Figs 1–3a) or 5% (other experiments). CO₂ was directly taken from the house-line and presented at 1%. Final flow rate of the air stream was set to 0.41 min⁻¹ (experiments in Figs 1–3a) or 1 l min⁻¹ (other experiments) with a final tubing size of 1/16 inch (inner diameter). Stability and reproducibility of the stimuli were continuously monitored throughout the experiments using a photo-ionization detector (PID; Aurora Scientific). A slightly different odour delivery system was used for electrophysiological experiments, in which air dilution was only one step and the odorants were diluted with mineral oil. Odour concentrations were adjusted to be equivalent to the 5% dilution used in imaging experiments, as confirmed by PID measurements. Final flow rate was 1 l min⁻¹. For all experiments, odours were presented in a pseudo-random order so that no odour was presented twice in succession.

Calcium imaging. *In vivo* two-photon calcium imaging was performed as described previously³ using a Prairie Ultima system (Bruker) and a Ti-Sapphire laser (Chameleon XR; Coherent) tuned to 920 nm (6–10 mW at the sample). All images were acquired with 40 \times water-immersion objective (LUMPlanFI/IR, numerical aperture 0.8; Olympus). The preparation was continuously perfused with saline containing (in mM): NaCl, 103; KCl, 3; CaCl₂, 1.5; MgCl₂, 4; NaHCO₃, 26; N-tris(hydroxymethyl)methyl-2-aminoethane-sulfonic acid, 5; NaH₂PO₄, 1; trehalose, 10; glucose, 10 (pH 7.3 when bubbled with 95% O₂ and 5% CO₂, 275 mOsm). When measuring odour tuning in MBONs, we aimed to separate signals from as many MBON types as possible, preferably at axons. However, many of the MBONs send their axons bilaterally in a symmetric manner. In these cases, even if the labelling in the split-GAL4 line is confined to a single cell in each brain hemisphere, it was often difficult to image a single axon because the symmetric axonal projections were extensively intertwined in a tight helical structure. Imaging at the dendrites was often advantageous to unambiguously assign signals to individual MBONs. Therefore we performed a series of pilot experiments to examine calcium responses in different cellular compartments. In the α 2sc neuron, we noticed that in some experiments there was almost no calcium response to any odour at the cell body (Extended Data Fig. 1a), while we had observed spiking responses to every odour in every whole-cell recording from this neuron (Fig. 3c). This disconnect is presumably attributable to low expression of voltage-gated

calcium channels at the soma and/or to the extremely long primary neurite connecting the soma to the axon and dendrites. On the other hand, calcium responses at dendrites and axons were consistently observed even when there was no somatic response (Extended Data Fig. 1a). Notably, the time course and the magnitude of the responses were largely similar between the dendrite and axon in the same neuron. We tested how general this was across different MBONs by comparing calcium responses in axons and dendrites in five different cell types. These five types were all the ones in which we could distinguish both axonal and dendritic signals for individual neurons (Extended Data Fig. 1a–e). In all cases, the responses were highly similar (Pearson's $r = 0.92 \pm 0.013$, mean \pm s.e.m.; $n = 10$ cells). Furthermore, tuning breadths were also identical (Extended Data Fig. 1f, g). Therefore, to measure tuning curves in MBONs, we imaged at either dendrite or axon, whichever maximized the isolation of signals from the MBON of interest (Extended Data Fig. 2; Extended Data Table 1), except for Extended Data Figs 1 and 8. KC imaging was performed at the cell body layer as described previously⁵. In both cases, imaging frames were typically 300 \times 300 pixels, acquired with a pixel dwell time of 1.6 μ s, yielding frame rates around 4 Hz, which slightly varied across experiments depending on the optical zoom factor. For each odour presentation trial, data were acquired for 20 s with an odour pulse (2 s in duration for experiments in Figs 1–3a, 1 s for other experiments) triggered 8 s after trial onset. Inter-stimulus interval was 25 s. When we compared tuning patterns of α 2sc neurons across hemispheres, we imaged the two hemispheres sequentially rather than simultaneously, so stimulus presentations were independent across the two recordings. This ensured that the higher correlations we observed within animals could not be attributed to noise correlations, that is, coordinated changes in neuronal responses based on momentary fluctuations in the internal state of the animal, such as its level of arousal³⁸.

Electrophysiology. Previously reported methods for *in vivo* whole-cell recordings in projection neurons¹⁹ were adapted for MBONs and KCs. The patch pipettes were pulled for a resistance of 4–5 M Ω (MBON) or 6–7 M Ω (KC) and filled with pipette solution containing (in mM): L-potassium aspartate, 125; HEPES, 10; EGTA, 1.1; CaCl₂, 0.1; Mg-ATP, 4; Na-GTP, 0.5; biocytin hydrazide, 13; with pH adjusted to 7.3 with KOH (265 mOsm). Bath solution was the same as in imaging experiments. Single or dual whole-cell current-clamp recordings were made using the Axon MultiClamp 700B amplifier (Molecular Devices). Cells were held at around -60 mV by injecting hyperpolarizing current (< 20 pA for MBONs, < 5 pA for KCs). Signals were low-pass filtered at 5 kHz and digitized at 10 kHz. Specific cell types were visually targeted by GFP signal with a 60 \times water-immersion objective (LUMPlanFI/IR; Olympus) attached to an upright microscope (BX51WI; Olympus). Although MZ160-GAL4 labels multiple types of MBONs in the MB-V2 cluster, we were typically able to distinguish the α 2sc neuron, which is a single unique neuron⁸, from the others based on the distinct size and location of its cell body. The morphology of all recorded cells, both α 2sc neurons and α/β KCs, were visualized by post-hoc immunohistochemistry with biocytin¹⁹; any data from incorrectly targeted cells were discarded. In dual whole-cell recording from α/β KCs and the α 2sc neuron, we took several steps to maximize the chance of detecting weak connections. Since KCs are immunonegative for choline acetyltransferase^{39,40} and unlikely to be cholinergic^{41,42}, we applied the cholinergic blocker mecamylamine (100 μ M) to the bath saline to minimize unrelated circuit activity that could obscure weak connections. In addition, we tested for connections using current injection (25.2 \pm 5.5 pA, 175 \pm 9.0 ms, mean \pm s.e.m.) to drive high frequency spike trains in the KCs (10.6 \pm 0.8 spikes, mean \pm s.e.m.).

Data analysis. All data analyses were performed in MATLAB (R2008b, MathWorks). All sample sizes were enough for robust statistical tests, which are appropriately chosen based on the distribution of data. No statistical methods were used to predetermine sample size.

Analysis of imaging data. For the analysis of MBONs, a region of interest (ROI) was manually set for each trial based on the mean baseline image. Response magnitudes were calculated as the integrated fluorescence change ($\Delta F/F$) in the time window between stimulus onset and 5 s after stimulus offset. Analysis of KC imaging required motion corrections and frame alignments in order to retain the identity of each ROI (that is, each KC cell body) throughout a whole imaging session, as described previously^{3,5}. For the measurement of tuning similarity, we used Pearson's correlation coefficient. When we used Euclidean distance instead, the results of the statistical tests remained the same in all cases (data not shown) except for one case (Extended Data Fig. 8e), where we did not detect a significant difference with the Euclidean distance measurement (data not shown).

Analysis of electrophysiological data. Spikes were automatically detected by custom-written scripts based on amplitude, after removing slow membrane potential deflections with a high-pass filter, and verified by visual inspection. Response spike rates were calculated in a window of 0.5–3.5 s after odour valve opening, and baseline spike rates were subtracted. To calculate the statistical significance of a connection between an α/β KC and an α 2sc neuron, we determined the difference

between the mean $\alpha 2sc$ membrane potential during the KC spike train and during the baseline period on each individual trial (51 ± 4 trials per pair), and used a t -test with a threshold significance value of $P < 0.05$. Out of 24 pairs recorded, we found only eight statistically significant postsynaptic responses. Five pairs were judged as monosynaptically connected because step-wise increments in membrane potential were obvious in spike-trigger-averaged traces and also because the delay between the KC spike and the onset of the EPSP was less than 2 ms (Extended Data Fig. 9a, d). In two other pairs, we observed smaller excitatory responses but could not detect unitary steps for each spike, even in spike-trigger-averaged traces (Extended Data Fig. 9b). Such connections could be either direct excitatory connections with unitary strength smaller than our detection limit or indirect feed-forward connections. The one remaining connection we found was, to our surprise, inhibitory (Extended Data Fig. 9c). The synaptic delay was too long for a monosynaptic connection via ionotropic transmission, suggesting that it represents a powerful feed-forward inhibitory connection. However, it remains possible that it is a monosynaptic connection with facilitatory synapses or slow metabotropic input. Although we cannot rule out the possibility that we are missing some extremely weak connections, we consider this unlikely given our experimental approach, namely quietening spontaneous network activity by blocking cholinergic transmission and then driving high frequency spike trains in the presynaptic KC (see the section above for details). Even if we assume that some connections went undetected, our results would nevertheless indicate that synaptic weights of KC– $\alpha 2sc$ connections are highly heterogeneous, since some connections were strong enough for us to detect clear unitary synaptic events. Thus, in any case our results indicate that connectivity between α/β KCs and $\alpha 2sc$ neurons is highly selective.

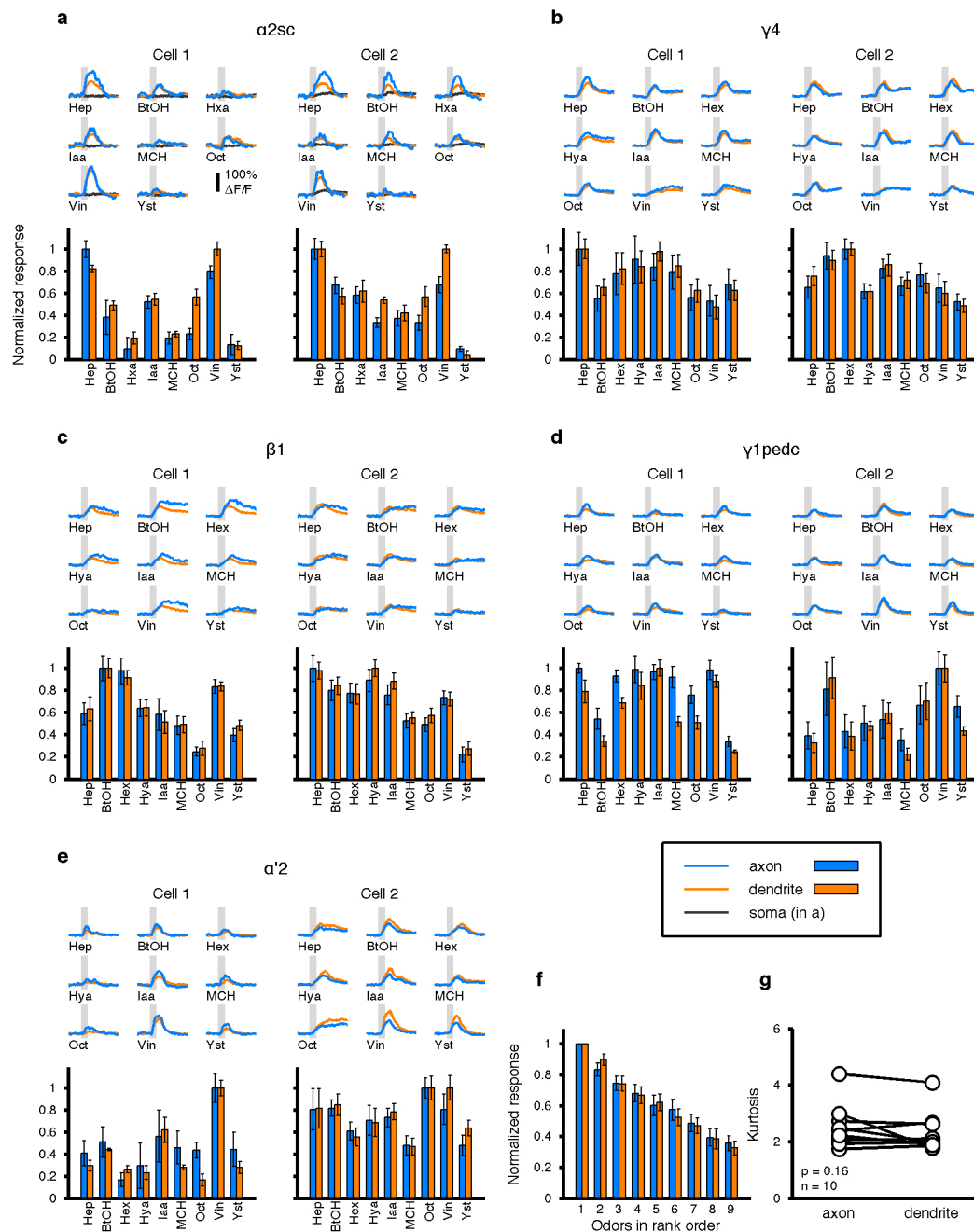
Population coding analyses. To compare population coding in KCs and MBONs, we represented the odour response patterns from each stimulus trial as a point in an N -dimensional neuronal coding space, where each of the axes represents the response magnitude of each neuron. For KC representations, each axis corresponded to one of the KCs imaged in one fly. For MBONs, since different cell types were imaged in different flies, we combined data from 17 different flies to produce an aggregate MBON population reflecting activity in a single 'virtual fly'. Since we have recordings from 5 flies for each type of MBON, by combining different recordings we can construct many different virtual flies; one example is shown in Fig. 2c. This approach assumes that there is no specific correlation between the tuning patterns of different cell types in the same animal. However, if different MBON types tended to be positively or negatively correlated in individual flies, that could have an impact on the odour-specific information available in population-level activity patterns, which we would have overlooked by analysing virtual flies. To test whether this was the case, we imaged multiple types of MBONs in the same animal by combining different split-GAL4 drivers (Extended Data Fig. 6). We compared tuning patterns in four pairs of cell types. These pairs included MBONs receiving input from the same MB lobe ($\alpha 1$ vs $\alpha 2sc$), those with input from the same types of KCs ($\alpha 1$ vs $\beta 1$), and those with input from different types of KCs ($\alpha 1$, $\beta 1$ vs $\gamma 4$), as well as MBONs with relatively higher variability of tuning across flies ($\alpha 1$ and $\alpha 2sc$; Fig. 3a). In no case did we find that the correlation of tuning within an animal was significantly different from the correlation across different animals (Extended Data Fig. 6). This result justifies using virtual flies to analyse MBON population representations. Each recording consisted of 4–7 trials per odour, so for virtual flies, we generated random combinations of those trials by bootstrapping. For example we generated 100 bootstrapped trials in one virtual fly for the display in Fig. 2c. On the other hand, we could not combine KC data from multiple flies because, unlike MBONs, the identities of KCs cannot be matched up between different flies. To make our analysis of the KC population similar to the MBONs, we also applied trial bootstrapping to the KC data, so that noise correlations, absent from the individually recorded MBON data, are also not present in the KCs.

Odour classification. For odour classification analysis, we adopted a linear classification algorithm based on Euclidean distances between odour representations in neuronal coding space. Odour-evoked activity patterns for each bootstrapped trial, generated as described above, were represented as points in this coding space, and the centroids of the points corresponding to each of the ten test odours were calculated. The data point for each trial was then classified as the odour with the nearest centroid. In order to avoid overfitting, the following two steps were implemented. First, when calculating centroid locations, the trial of interest was removed from the data set (leave-one-out cross validation). Second, when generating bootstrapped trials, the same trial was selected only once in a given data set. Since the number of trials in each recording was 4–7, we generated only 4–7 bootstrapped trials at a time. For KCs, we repeated this process 10 times per fly ($n = 10$ flies), yielding 100 classification scores in Fig. 2d. For MBONs, we generated one combination of bootstrapped trials for each virtual fly, and repeated this in 100 different virtual flies, again yielding 100 data points for Fig. 2d. To artificially decorrelate tuning profiles across MBONs, we shuffled odour labels for

each cell's data, which breaks up any tendency for odours to evoke similar responses in different MBONs. These newly assigned odour labels were fixed and used for the subsequent classification analysis with same nearest-centroid approach described above. This whole process, shuffling odours, determining centroids and then testing classification, was carried out 100 separate times to obtain the 100 classification scores plotted in Fig. 2d. When neuron labels are shuffled between trials, classification accuracy dropped to near chance level, indicating that patterns of activity are important for classification in both cases. They did not drop all the way to chance because different odours evoke different overall levels of activity, which also contributes information about odour identity; this is eliminated by additionally shuffling odour labels (Fig. 2d).

Cluster analysis. For cluster analysis (Fig. 2f), we employed the k -means clustering algorithm. k -means clustering is an unsupervised clustering method that allows us to partition data points into arbitrary number of clusters such that the total distance from the individual points to their cluster centroids is minimized. The quality of such clusters can be evaluated by the mean silhouette value of all the data points⁴³. The silhouette value (ranging from -1 to 1) for the i -th data point can be calculated as $s(i) = (b(i) - a(i))/\max\{a(i), b(i)\}$, where $a(i)$ is the average distance to the other points in the same cluster, and $b(i)$ is the average distance to the points in the closest neighbouring cluster. Thus, if data points form compact clusters that are well separated from each other and are appropriately partitioned, the mean silhouette value approaches 1 . On the other hand, if clusters are diffuse and/or overlapping each other, a lower value is obtained, even when the clusters are optimally partitioned. In addition, even when there are nicely separated compact clusters, it gives a lower value if the partitioning pattern is inappropriate (for example, dividing one cluster into halves or combining two clusters to one). Thus, by using k -means to explore a wide-ranging number of clusters (2 to 20), and searching for the partitioning that gave the highest silhouette value, we were able to determine the optimal number of clusters in MBON and KC odour representations. We examined the clustering in the data with different dimensionalities by gradually increasing the number of principal components (PCs) for the projection. To make the analysis comparable between KCs and MBONs, which consist of profoundly different numbers of cells, we matched the total fraction of variance captured as we increased dimensionality, rather than directly increasing the number of PCs. In other words, we increased dimensionality by using a sufficient number of PCs to capture the same amount of variance in KC and MBON populations. Cluster quality quickly diminished when too many PCs were added. This is presumably because these later PCs mainly contained noise rather than signals related to odours, which would simply make existing clusters more diffuse in higher-dimensional space. We used built-in functions of MATLAB to perform PCA and k -means clustering as well as to calculate silhouette values.

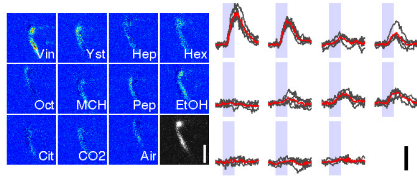
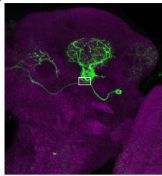
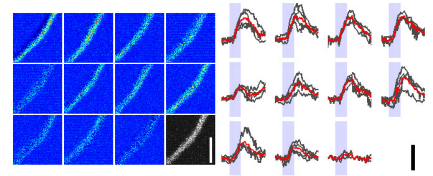
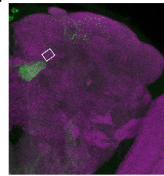
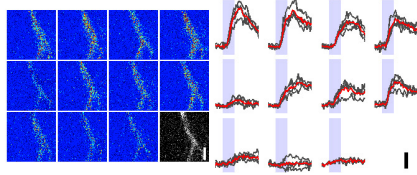
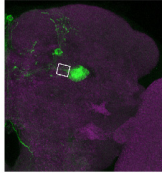
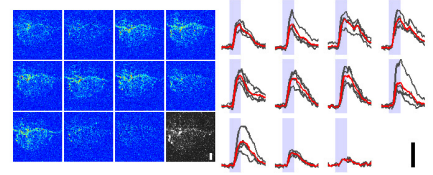
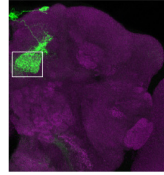
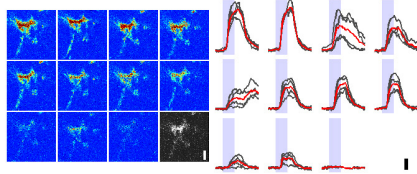
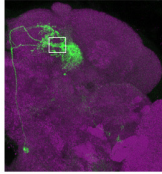
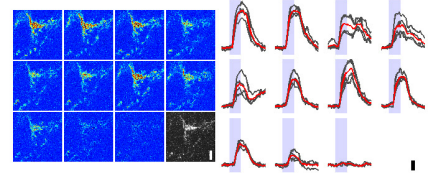
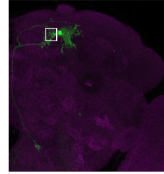
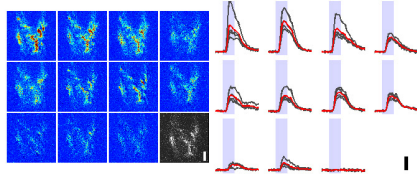
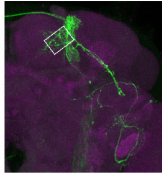
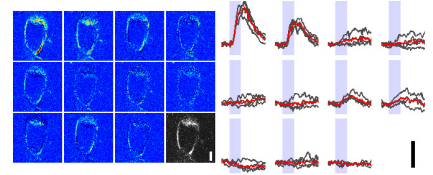
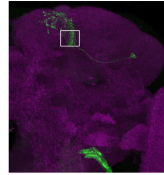
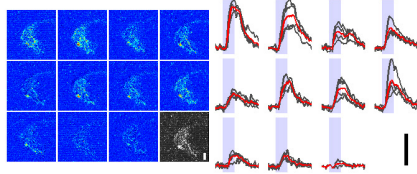
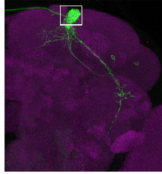
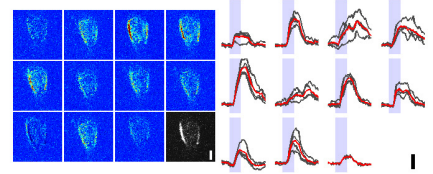
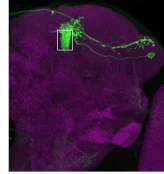
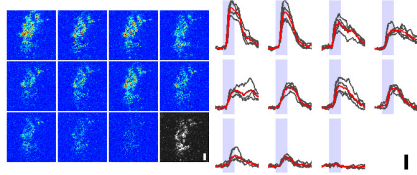
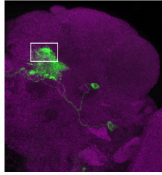
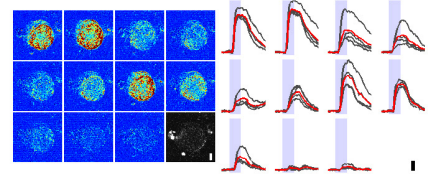
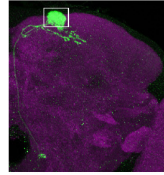
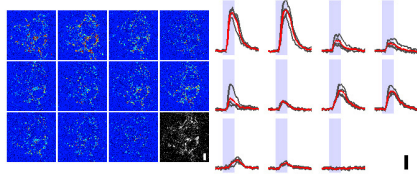
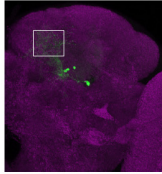
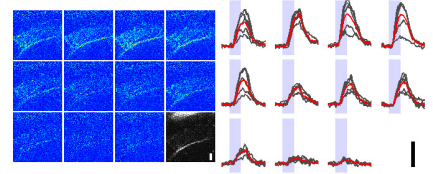
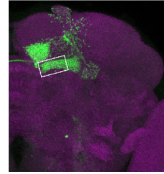
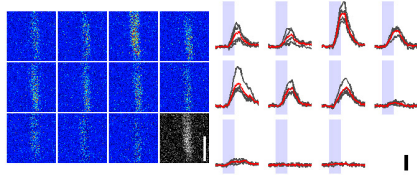
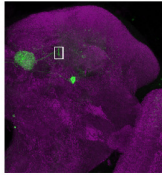
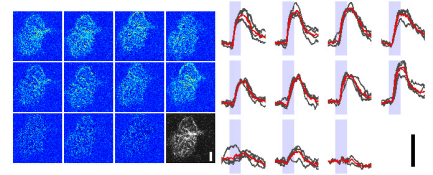
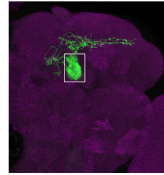
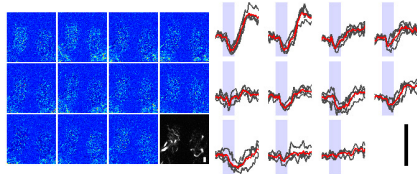
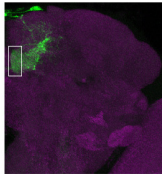
- Chen, T.-W. *et al.* Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499**, 295–300 (2013).
- Akerboom, J. *et al.* Optimization of a GCaMP calcium indicator for neural activity imaging. *J. Neurosci.* **32**, 13819–13840 (2012).
- Connolly, J. B. *et al.* Associative learning disrupted by impaired G_s signaling in *Drosophila* mushroom bodies. *Science* **274**, 2104–2107 (1996).
- Aso, Y. *et al.* The mushroom body of adult *Drosophila* characterized by GAL4 drivers. *J. Neurogenet.* **23**, 156–172 (2009).
- Han, P. L., Levin, L. R., Reed, R. R. & Davis, R. L. Preferential expression of the *Drosophila rutabaga* gene in mushroom bodies, neural centers for learning in insects. *Neuron* **9**, 619–627 (1992).
- Ito, K. *et al.* The organization of extrinsic neurons and their implications in the functional roles of the mushroom bodies in *Drosophila melanogaster* Meigen. *Learn. Mem.* **5**, 52–77 (1998).
- Yang, M. Y., Armstrong, J. D., Vilinsky, I., Strausfeld, N. J. & Kaiser, K. Subdivision of the *Drosophila* mushroom bodies by enhancer-trap expression patterns. *Neuron* **15**, 45–54 (1995).
- Averbeck, B. B., Latham, P. E. & Pouget, A. Neural correlations, population coding and computation. *Nature Rev. Neurosci.* **7**, 358–366 (2006).
- Gorczyca, M. G. & Hall, J. C. Immunohistochemical localization of choline acetyltransferase during development and in *ChA*^{ts} mutants of *Drosophila melanogaster*. *J. Neurosci.* **7**, 1361–1369 (1987).
- Yasuyama, K., Meinertzhagen, I. A. & Schürmann, F.-W. Synaptic organization of the mushroom body calyx in *Drosophila melanogaster*. *J. Comp. Neurol.* **445**, 211–226 (2002).
- Johard, H. A. D. *et al.* Intrinsic neurons of *Drosophila* mushroom bodies express short neuropeptide F: relations to extrinsic neurons expressing different neurotransmitters. *J. Comp. Neurol.* **507**, 1479–1496 (2008).
- Brooks, E. S. *et al.* A putative vesicular transporter expressed in *Drosophila* mushroom bodies that mediates sexual behavior may define a neurotransmitter system. *Neuron* **72**, 316–329 (2011).
- Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
- Willmore, B. & Tolhurst, D. J. Characterizing the sparseness of neural codes. *Network* **12**, 255–270 (2001).
- Chiang, A.-S. *et al.* Three-dimensional reconstruction of brain-wide wiring networks in *Drosophila* at single-cell resolution. *Curr. Biol.* **21**, 1–11 (2011).



Extended Data Figure 1 | Comparison of odour-evoked calcium responses across multiple cell compartments. **a**, Odour responses were sequentially recorded at three different compartments (axon, blue; dendrite, orange and soma, black) in two different $\alpha 2sc$ neurons with GCaMP5 imaging. Upper subpanel shows mean $\Delta F/F$ traces from different compartments (4 or 5 trials for each). Shaded areas indicate 1-s odour stimulations. Lower subpanel shows tuning profiles in axon and dendrites (normalized to the strongest response, mean \pm s.e.m.). Note that the time course, response magnitude and tuning profiles are largely consistent between the axon and dendrite in the same cell,

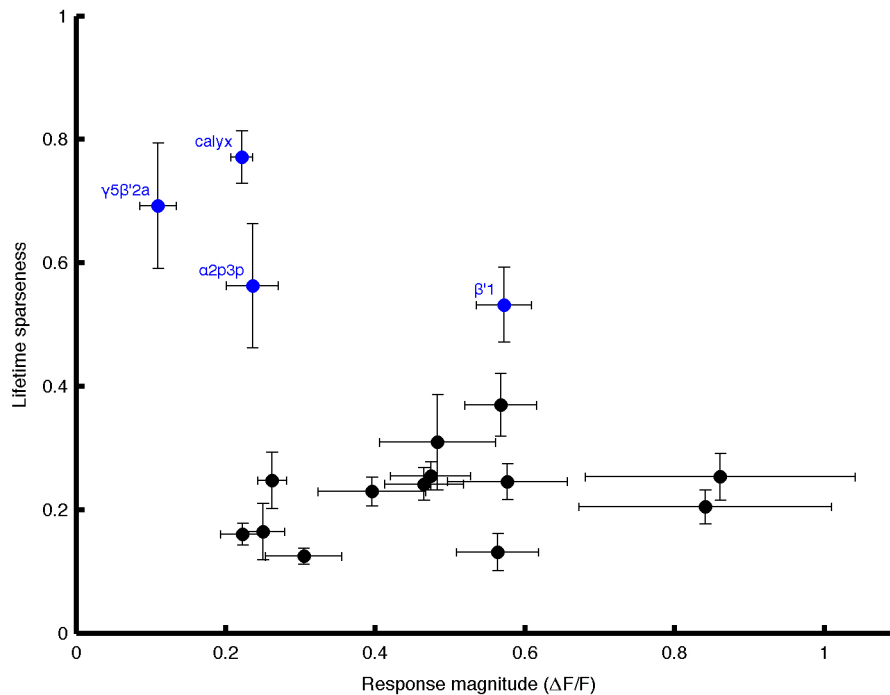
while signals at the soma are small and slower (Cell2) or sometimes undetectable (Cell1). **b–e**, Data from 4 other types of MBONs. Scale of the traces is the same as in **a**. Correlation between axons and dendrites is 0.92 ± 0.013 (Pearson's r , mean \pm s.e.m.; $n = 10$ cells). Hya, hexyl acetate. **f**, Mean normalized tuning of the ten cells. Odors are sorted according to the rank order in each cell to visualize the tuning width. **g**, Kurtosis of tuning curves. There was no difference between axons and dendrites ($P = 0.16$, paired t -test).

calyx

 $\beta 2\beta'2a$  $\gamma 1pedc$  $\beta'2mp$  $\gamma 2\alpha'1$  $\alpha'2$  $\alpha'1$  $\alpha 2p3p$  $\alpha'3ap \& m$  $\alpha 2sc$  $\gamma 3 \& \gamma 3\beta'1$  $\alpha 3$  $\beta'1$  $\beta 1$  $\gamma 4$  $\alpha 1$  $\gamma 5\beta'2a$ 

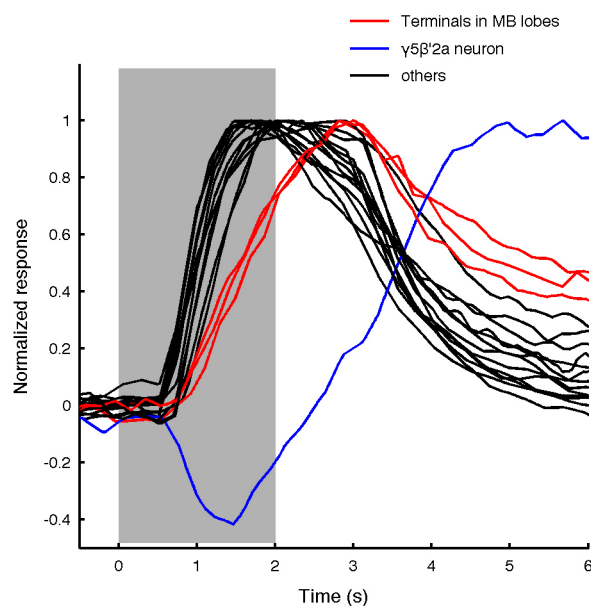
$\Delta F/F$ 0 1 2 3 4 >5

Extended Data Figure 2 | Raw traces of calcium imaging in MBONs. For all 17 types/combination of types of MBONs, we show a projection image of confocal stacks from split-GAL4 lines (left), example $\Delta F/F$ images of the calcium responses to ten odours and air (middle) and $\Delta F/F$ time courses (right). Note that all the split-GAL4 lines used in this study label the target MBONs with extremely high specificity. The white squares indicate the approximate region scanned during calcium imaging. Greyscale images show baseline fluorescence (scale bar, 5 μm), while colour map images represent calcium responses ($\Delta F/F$, colour scale bottom right). Traces showing $\Delta F/F$ time courses on individual trials (grey; $n = 4-7$) are overlaid with the mean (red), scale bar, 50% $\Delta F/F$.



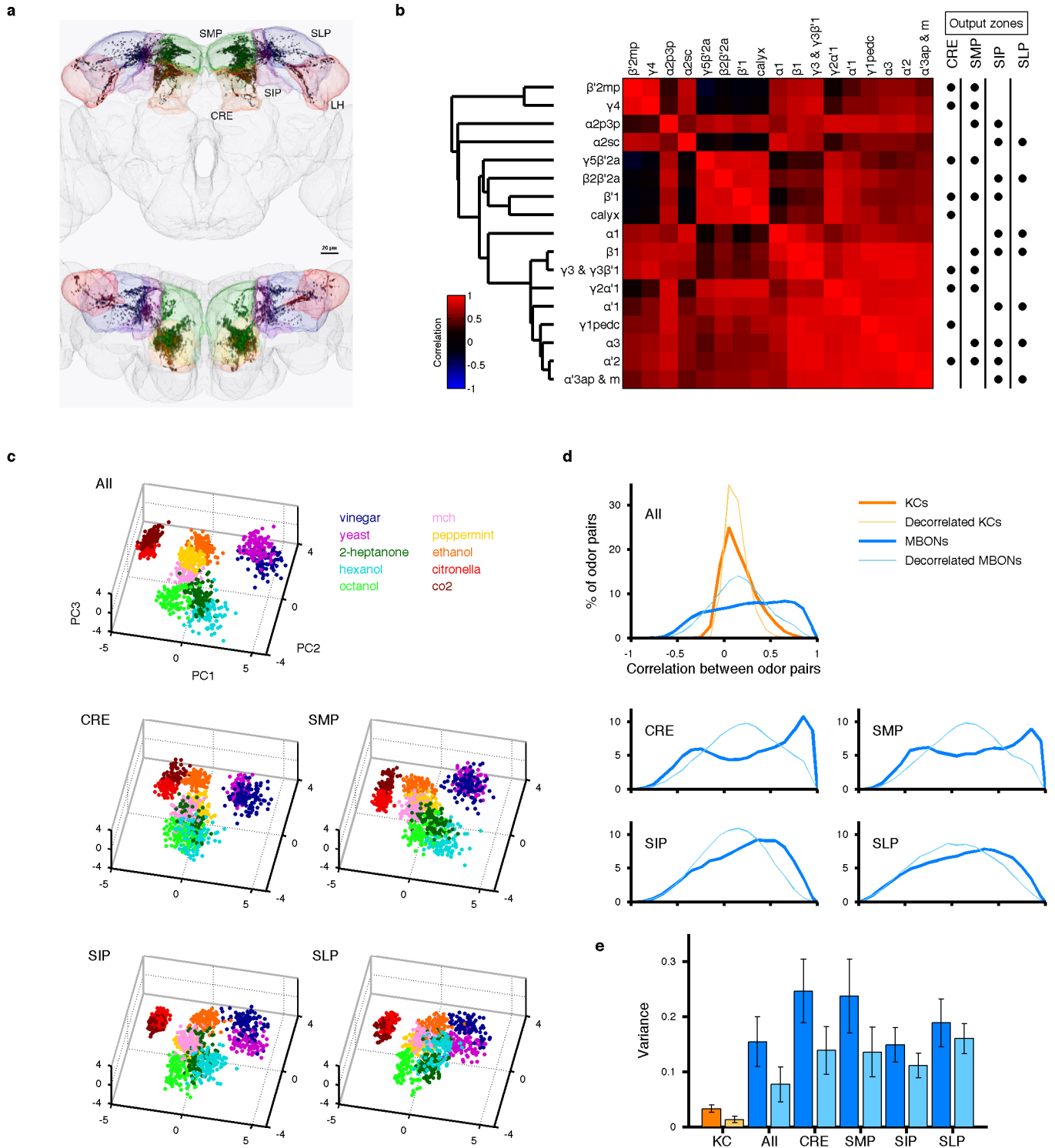
Extended Data Figure 3 | Relationship between response intensity and tuning selectivity. Lifetime sparseness, an index of tuning selectivity^{1,44}, was plotted against the magnitude of the largest odour response observed for each neuron (mean $\Delta F/F$ during the response window). Each dot corresponds to one of the 17 MBON types (mean \pm s.e.m.; $n = 5$ flies). There was no correlation between the two variables (Spearman's $\rho = -0.16$, $P = 0.54$). Thus, the relatively selective tuning patterns observed in four MBONs (blue) are likely not related to the intensity of the response. Rather, for three of these cells, we noticed that their narrow tuning was accompanied by unusual

dendritic anatomy. $\beta'1$ neuron is the only MBON with extensive dendritic processes outside the MB, and calyx neuron (MB-CP1) is the only MBON that samples from the MB calyx. $\alpha 2p3p$ neuron's dendrites seem to contact solely the $\alpha/\beta p$ KCs, whose dendrites arborize in the sub-region of the MB calyx known as the accessory calyx, where no olfactory input has been reported⁷. The narrow tuning patterns of the fourth neuron, $\gamma 5\beta'2a$, seem to be related to the transient inhibitory epochs uniquely observed in this neuron (Extended Data Fig. 4).



Extended Data Figure 4 | Diversity in response time courses across the MBON population. Normalized mean $\Delta F/F$ responses to yeast odour for all MBON types overlaid (shading indicates timing of odour presentation).

Note that three MBONs ($\beta 1$, $\gamma 1$ pedc, and $\gamma 4$) that have axonal terminals in the MB lobes show slower time courses (red) than the others (black). The cell with the characteristic inhibitory period is the $\gamma 5\beta'2a$ neuron (blue).



Extended Data Figure 5 | Population analysis of subpopulations of MBONs.

a. Distribution of MBON terminals in the brain (dark colours), based on the localization of a presynaptic marker⁸ (Syt::smGFP-HA). Anterior (top) and dorsal (bottom) views are shown. Axonal projections of MBONs converge heavily onto small areas inside four neighbouring neuropils (light colours), crepine (CRE; orange), superior medial protocerebrum (SMP; green), superior intermediate protocerebrum (SIP; purple) and superior lateral protocerebrum (SLP; blue). We have found numerous cells that send widely branching neurites into one or all of these areas (database associated with ref. 45), indicating that downstream neurons likely read out activity from populations of MBONs. Lateral horn (LH, red) also receives sparse input.

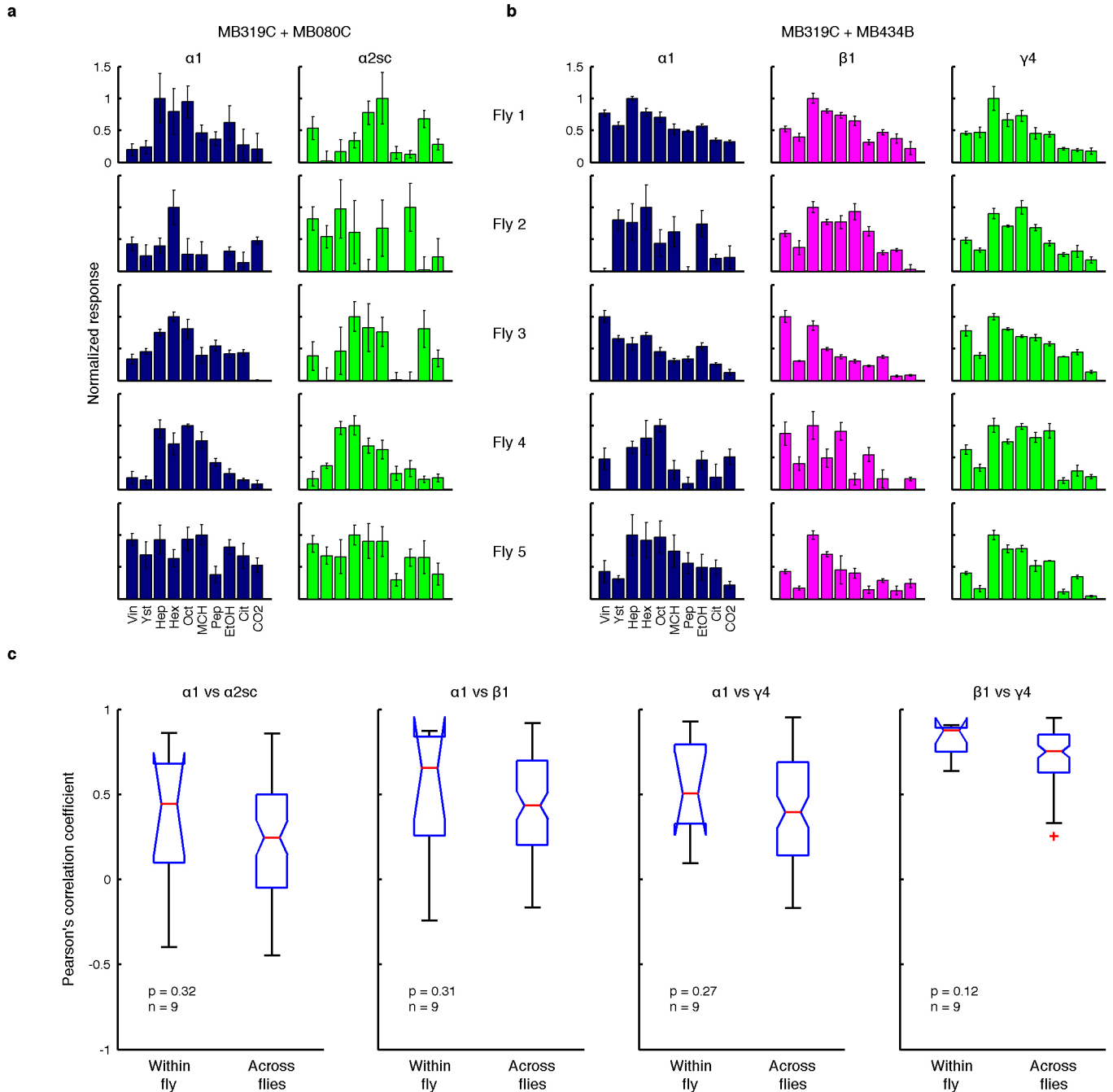
b. Pairwise correlations of MBON tuning patterns and the corresponding dendrogram are shown in the same way as Fig. 1d. The dots on the right show the axonal projection site of the different MBON types. None of the four major projection zones samples from particularly decorrelated set of MBONs.

c. Odour representations in MBONs from a single virtual fly visualized by PCA

(different virtual fly from Fig. 2c). Representations from the full MBON population (All) and subpopulations with the same axonal projection zones (CRE, SMP, SIP and SLP) are shown. Representations by subpopulations tend to be noisier than those from the full population, but they retain grossly similar structures.

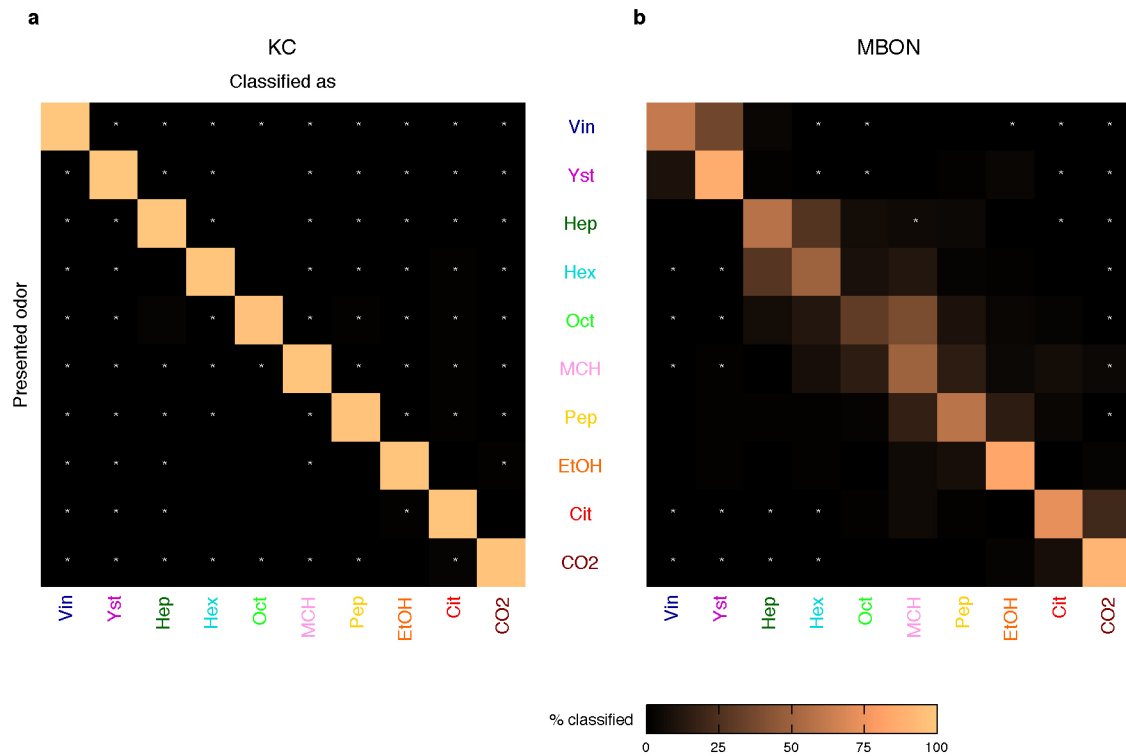
d. Correlation coefficients of neural representations of ten odours in KCs and MBONs (thick lines), as in Fig. 2g, h. Correlation values observed after artificially decorrelating KC and MBON tuning curves are shown for comparison (thin lines).

e. Variance of the correlation coefficients shown in **d** (dark bars; mean \pm s.d.; $n = 1,000$ virtual flies), showing that values are more widely distributed in MBONs compared to KCs. Artificially decorrelating MBON tuning curves reduces the variance considerably (light bars), indicating that the pattern of activity in the MBONs contributes to the high variance. However, variance remains greater than that in the KCs, indicating that the breadth of tuning in the MBONs also contributes to wide range of correlation coefficients. The ratio of the contribution of those two factors seems to be different across subpopulations projecting different zones.



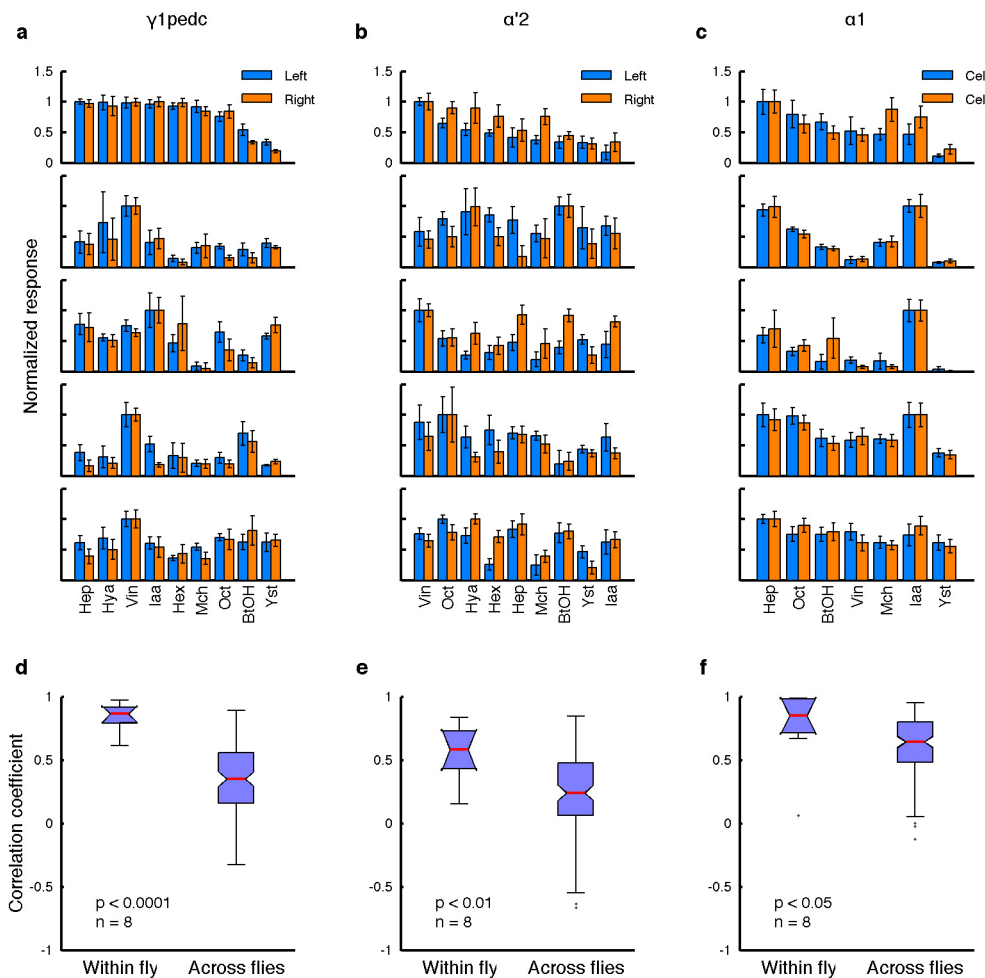
Extended Data Figure 6 | Correlation of tuning between multiple MBON types measured in the same animal. **a**, Odour tuning of $\alpha 1$ and $\alpha 2sc$ neurons recorded sequentially in the same fly with GCaMP5 imaging (normalized to the strongest response, mean \pm s.e.m.). Data from five representative flies are shown. Two split-GAL4 drivers, MB319C and MB080C, were combined to label the two cell types. **b**, Similar to **a** but recordings are from $\alpha 1$, $\beta 1$ and $\gamma 4$ neurons. MB319C was combined with MB434B that labels both $\beta 1$ and $\gamma 4$

neurons. **c**, Correlation of tuning between different cell types, calculated within and across flies. Four different pairwise combinations of cell types were examined. Although within-fly correlations tended to be slightly higher than across-fly correlations, none of the four cases was significant (Mann-Whitney *U*-test). This is in sharp contrast with the comparisons of the same cell type, where we observed much higher correlations within the same fly than across flies (Fig. 3).



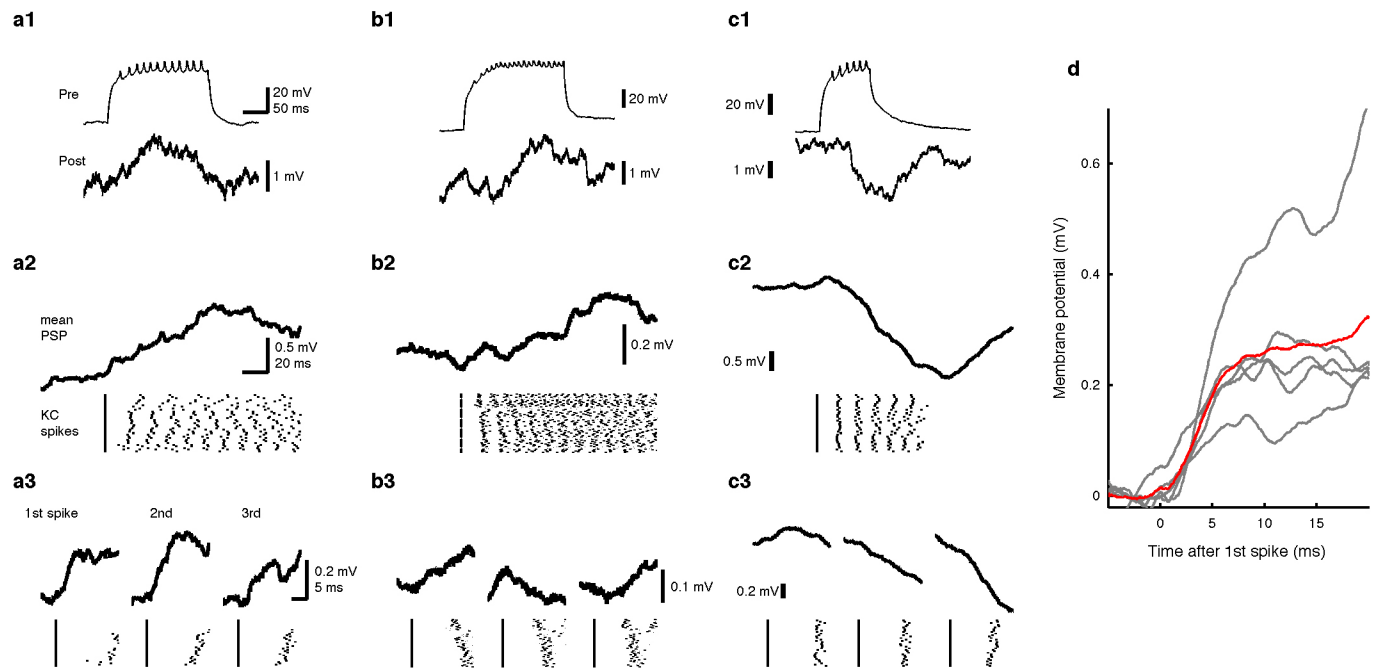
Extended Data Figure 7 | Confusion matrices from odour classification analysis. Confusion matrices generated from the classification analysis in Fig. 2d. **a**, KCs show nearly perfect classification performance. **b**, MBONs fail to discriminate some odours. Although the confusion matrix was generated using 100 different virtual flies (see Methods), a high rate of misclassifications is

observed only between certain odour pairs that form overlapping clouds in MBON space (Fig. 2c). Odour pairs that were never misclassified with each other are indicated with white asterisks. Note that the group of vinegar and yeast were never misclassified with the group of CO₂ and citronella.



Extended Data Figure 8 | Individualized tuning in multiple different MBON types. **a–c**, Odour tuning of pairs of MBONs in the same fly, from GCaMP5 imaging at axons (normalized to the strongest response, mean \pm s.e.m.). Data from five representative flies are shown. For $\gamma 1pedc$ (**a**) and $\alpha'2$ neurons (**b**), cells on the left and right hemispheres are compared.

For $\alpha 1$ neuron (**c**), recordings were from ipsilateral cell pairs. **d–f**, For all three cell types, tuning patterns of the neurons from the same fly are more correlated than those from different flies ($n = 8$ flies per cell types; Mann–Whitney U -test).



Extended Data Figure 9 | Diverse functional connections between KCs and a MBON. **a**, Another example of a monosynaptically connected pair of α/β KC and the $\alpha 2sc$ neuron from experiments shown in Fig. 4. **a1**, Sample traces of simultaneously recorded KC (Pre) and the $\alpha 2sc$ neuron (Post). **a2**, The $\alpha 2sc$ neuron's membrane potential spike-trigger-averaged on the first KC spike of each trial. Raster plot of the KC spikes is shown below. **a3**, Enlarged spike-trigger-averaged EPSPs for the first, second and third spikes in the train. **b**,

A pair with small excitatory connection that we could not confirm as monosynaptic. In total, two of such pairs were found. **c**, A pair with an inhibitory connection, which is likely to be polysynaptic. Only one such pair was found. **d**, Summary of all five monosynaptically connected pairs. The spike-trigger-averaged EPSPs from the first spike in the train from five different recordings are shown overlaid in grey. The mean across cells is shown in red.

Extended Data Table 1 | MBON types and imaging conditions.

| Type* (Dendritic site in MB) | Axonal terminals in MB* | No. of cells per hemisphere* | Imaged strain* | Signal isolation | Imaged locus |
|---------------------------------|----------------------------|---------------------------------|-------------------|---------------------|---------------------|
| $\alpha 1$ | - | 2 | MB310C | b | dendrite |
| $\alpha 2p3p$ | - | 2 | MB062B | b | dendrite |
| $\alpha 2sc$ | - | 1 | MB080C | a | dendrite |
| $\alpha 3$ | - | 2 | MB093C | b | dendrite |
| $\beta 1$ | α | 1 | MB434B | a | dendrite |
| $\beta 2\beta'2a$ | - | 1 | MB399B | a | dendrite shaft |
| $\alpha'1$ | - | 2 | MB543B | b | axon terminals |
| $\alpha'2$ | - | 1 | MB018B | a | axon arbor |
| $\alpha'3ap$ | - | 1 | MB027B | c | dendrite |
| $\alpha'3m$ | - | 2 | | | |
| $\beta'1$ | - | 8 | MB057B | b | axon arbor |
| $\beta'2mp$ | - | 1 | MB002B | a | dendrite |
| $\beta'2mp_bilateral$ | - | 1 | N.A. | N.A. | N.A. |
| $\gamma 1 pedc$ | α/β | 1 | MB085C | a | dendrite shaft |
| $\gamma 1\gamma 2$ | - | 1 | N.A. | N.A. | N.A. |
| $\gamma 2\alpha'1$ | - | 2 | MB051B | b | axon arbor |
| $\gamma 3$ | - | 1 | MB083C | c | axon terminals |
| $\gamma 3\beta'1$ | - | 1 | | | |
| $\gamma 4$ | $\gamma 1\gamma 2$ | 1 | MB298B | a | axon shaft |
| $\gamma 4\gamma 5$ | - | 1 | N.A. | N.A. | N.A. |
| $\gamma 5\beta'2a$ | - | 1 | MB011B | a | $\gamma 5$ dendrite |
| calyx (MB-CP1) | - | 1 | MB242A | a | dendrite shaft |

^aImaging signal is from single neuron. ^bSignal is from multiple cells of single cell type. ^cSignal is from multiple cell types. *Data from ref. 8.

The genomic landscape of response to EGFR blockade in colorectal cancer

Andrea Bertotti^{1,2,3*}, Eniko Papp^{4*}, Siân Jones⁵, Vilmos Adleff⁴, Valsamo Anagnostou⁴, Barbara Lupo^{1,2}, Mark Sausen⁵, Jillian Phallen⁴, Carolyn A. Hruban⁴, Collin Tokheim⁶, Noushin Niknafs⁶, Monica Nesselbush⁵, Karli Lytle⁵, Francesco Sassi², Francesca Cottino², Giorgia Migliardi^{1,2}, Eugenia R. Zanella^{1,2}, Dario Ribero^{7†}, Nadia Russolillo⁷, Alfredo Mellano², Andrea Muratore², Gianluca Paraluppi⁸, Mauro Salizzoni^{8,9}, Silvia Marsoni², Michael Kragh¹⁰, Johan Lantto¹⁰, Andrea Cassingena¹¹, Qing Kay Li⁴, Rachel Karchin^{4,6}, Robert Scharpf⁴, Andrea Sartore-Bianchi¹¹, Salvatore Siena^{11,12}, Luis A. Diaz Jr^{4,13}, Livio Trusolino^{1,2§} & Victor E. Velculescu^{4§}

Colorectal cancer is the third most common cancer worldwide, with 1.2 million patients diagnosed annually. In late-stage colorectal cancer, the most commonly used targeted therapies are the monoclonal antibodies cetuximab and panitumumab, which prevent epidermal growth factor receptor (EGFR) activation¹. Recent studies have identified alterations in *KRAS*^{2–4} and other genes^{5–13} as likely mechanisms of primary and secondary resistance to anti-EGFR antibody therapy. Despite these efforts, additional mechanisms of resistance to EGFR blockade are thought to be present in colorectal cancer and little is known about determinants of sensitivity to this therapy. To examine the effect of somatic genetic changes in colorectal cancer on response to anti-EGFR antibody therapy, here we perform complete exome sequence and copy number analyses of 129 patient-derived tumour grafts and targeted genomic analyses of 55 patient tumours, all of which were *KRAS* wild-type. We analysed the response of tumours to anti-EGFR antibody blockade in tumour graft models and in clinical settings and functionally linked therapeutic responses to mutational data. In addition to previously identified genes, we detected mutations in *ERBB2*, *EGFR*, *FGFR1*, *PDGFRA*, and *MAP2K1* as potential mechanisms of primary resistance to this therapy. Novel alterations in the ectodomain of *EGFR* were identified in patients with acquired resistance to EGFR blockade. Amplifications and sequence changes in the tyrosine kinase receptor adaptor gene *IRS2* were identified in tumours with increased sensitivity to anti-EGFR therapy. Therapeutic resistance to EGFR blockade could be overcome in tumour graft models through combinatorial therapies targeting actionable genes. These analyses provide a systematic approach to evaluating response to targeted therapies in human cancer, highlight new mechanisms of responsiveness to anti-EGFR therapies, and delineate new avenues for intervention in managing colorectal cancer.

To examine genetic alterations that affect response to anti-EGFR therapy, we selected 137 colorectal cancers (CRCs) from liver metastases that were *KRAS* wild-type as determined by Sanger sequencing (Supplementary Table 1). To elucidate genetic alterations in these cancers, we enriched for neoplastic cells using patient-derived tumour grafts and performed exome sequencing of tumour graft and matched normal DNA (Supplementary Tables 1 and 2). This approach identified sequence changes and copy number alterations in more than 20,000

genes, with an average coverage within the target regions of nearly 150-fold for each sample (Supplementary Tables 3 and 4).

Sequence analyses of 135 of 137 tumours identified a median of 117 somatic mutations in each cancer. Two tumours displayed an elevated number of somatic alterations (2,979 and 2,480 changes per exome), consistent with a mutator phenotype. Common CRC driver genes were identified at expected frequencies in the tumours analysed (Supplementary Tables 3–5). Eight tumours were identified as having *KRAS* alterations that were not initially detected by Sanger sequencing and were excluded from further analysis, resulting in 129 *KRAS* wild-type tumours.

To evaluate whether identified alterations were associated with resistance to EGFR inhibitors, we determined tumour graft response to cetuximab therapy for 116 of the 129 *KRAS* wild-type CRCs (Figs 1 and 2). The volume of each tumour graft was evaluated at 3 and 6 weeks, and tumours were categorized as showing disease progression, regression, or stabilization. Among tumour grafts with disease progression (increase in tumour volumes over 35%) or suboptimal stabilization (increase in tumour volumes between 20 and 35%), we detected alterations in all genes thought to be involved in EGFR therapeutic resistance: *NRAS* codon 12 or 61 mutations (seven cases), *BRAF* V600E mutation (three cases), *MET* amplification (three cases), and *ERBB2* amplification (four of five cases). Additionally, three out of four tumours with alterations in exon 20 of *PIK3CA* and four out of five tumours with protein truncating or homozygous deletions of *PTEN* were resistant to anti-EGFR blockade.

We evaluated potential mechanisms of resistance that have not been previously described in CRC. We focused on cell-surface receptors or members of the EGFR signalling pathway to identify candidate genes that were altered in therapy-resistant tumours (Fig. 2, Extended Data Fig. 1 and Supplementary Tables 3 and 4). We observed point mutations affecting the *ERBB2* kinase domain, including in two tumours with the same change at V777L and another tumour harbouring an L866M mutation, as well as a sequence change in the ectodomain at S310Y, all of which correlated with cetuximab resistance. Although amplification of *ERBB2* has been reported in CRCs^{9,10,14}, sequence alterations in this gene have not been linked to therapeutic resistance to anti-EGFR blockade. These data suggest that somatic mutations in *ERBB2* may provide an alternative mechanism for *ERBB2* pathway activation that is complementary to *ERBB2* amplification in CRC. Similarly, we found sequence alteration in the kinase domain of

¹Department of Oncology, University of Turin Medical School, 10060 Candiolo, Turin, Italy. ²Translational Cancer Medicine, Surgical Oncology, and Clinical Trials Coordination, Candiolo Cancer Institute – Fondazione del Piemonte per l'Oncologia IRCCS, 10060 Candiolo, Turin, Italy. ³National Institute of Biostructures and Biosystems (INBB), 00136 Rome, Italy. ⁴Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, Maryland 21287, USA. ⁵Personal Genome Diagnostics, Baltimore, Maryland 21224, USA. ⁶Department of Biomedical Engineering, Institute for Computational Medicine, Johns Hopkins University, Baltimore, Maryland 21204, USA. ⁷Department of Surgery, Maurizio Umberto I Hospital, 10128 Turin, Italy. ⁸Liver Transplantation Center, San Giovanni Battista Hospital, 10126 Turin, Italy. ⁹Department of Surgical Sciences, University of Turin Medical School, 10126 Turin, Italy. ¹⁰Symphogen A/S, 2750 Ballerup, Denmark. ¹¹Niguarda Cancer Center, Ospedale Niguarda Ca' Granda, 20162 Milan, Italy. ¹²University of Milan Medical School, 20162 Milan, Italy. ¹³Swim Across America Laboratory, The Ludwig Center for Cancer Genetics and Therapeutics at Johns Hopkins, Baltimore, Maryland 21287, USA. [†]Present address: European Institute of Oncology (IEO), 20141 Milan, Italy.

*These authors contributed equally to this work.

§These authors jointly supervised this work.

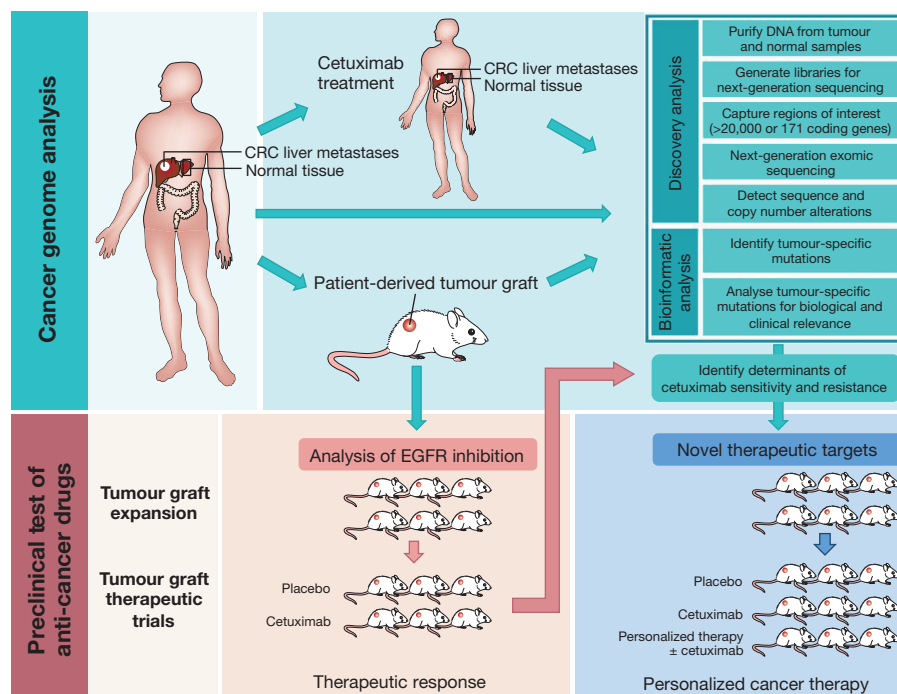


Figure 1 | Schematic diagram of integrated genomic and therapeutic analyses. To examine the effect of genomic alterations on sensitivity to anti-EGFR blockade, we performed whole-exome and copy-number analyses of 129 early-passage tumour grafts and targeted analyses of 55 tumours from patients, all of which were *KRAS* wild-type (top). Twenty-two of the tumour grafts were from patients who had been previously treated with anti-EGFR therapy.

One hundred and sixteen of these tumour grafts were evaluated for response to cetuximab in preclinical therapeutic trials (bottom left). Integration of genomic and therapeutic information was used to identify candidate resistance and response genes, and to design preclinical trials using novel compounds to overcome resistance to EGFR blockade (bottom right).

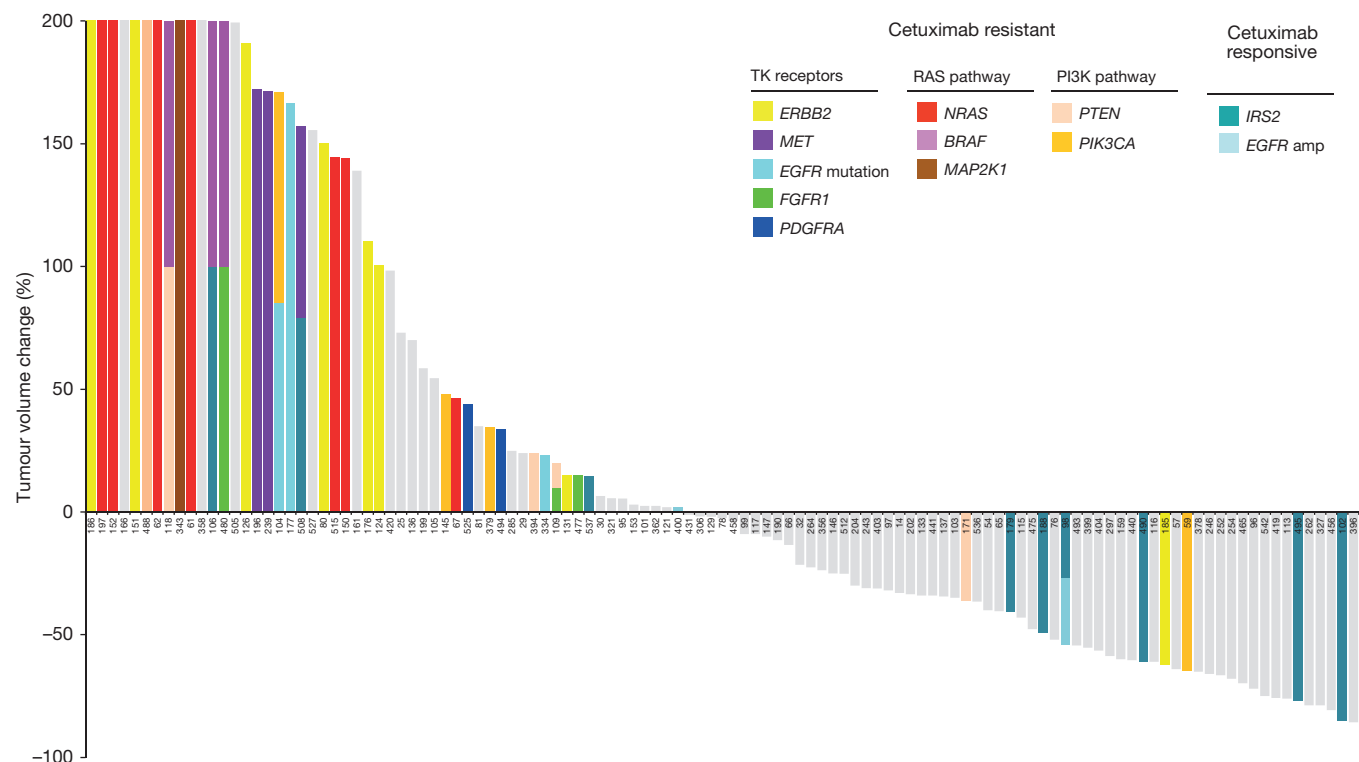


Figure 2 | Effect of cetuximab treatment on growth of colorectal tumours with different somatic alterations. Waterfall plot of tumour volume changes after cetuximab treatment, compared with baseline, in 116 *KRAS* wild-type tumour grafts. Alterations related to therapeutic resistance or sensitivity are shown in the indicated colours (complete lists of alterations are in Supplementary Tables 3, 4 and 6). For the following genes, a subset of

alterations is indicated: *MET* amplification; *FGFR1* amplification; *PDGFRA* kinase domain mutations; *BRAF* V600 hotspot mutations; *PTEN* homozygous deletion or truncating mutations; *PIK3CA* exon 20 mutations; *EGFR* ecto- and kinase domain mutations and amplifications. The maximum threshold for tumour growth was set at 200%.

EGFR (V843I) in one case that showed tumour growth in the presence of cetuximab. Although *EGFR* kinase alterations are rare in CRC^{15,16}, the observed case suggests that in principle such changes may provide a mechanism of resistance to anti-*EGFR* therapy.

We identified alterations in additional protein kinase receptors in tumours resistant to cetuximab treatment: amplification of the fibroblast growth factor receptor *FGFR1* and sequence alterations in the platelet-derived growth factor receptor *PDGFRA*. Each of these was altered in four of the 129 CRC samples analysed (8 samples in total, 6%). *FGFR1* is a known driver in human cancers¹⁷ and has been reported to be amplified in different tumour types. *PDGFRA* is a tyrosine kinase receptor that is known to be mutated in gastrointestinal stromal tumours¹⁸. The detected sequence alterations in *PDGFRA*, including a mutation that affected the same residue in two different patients (R981H), were all located in or near the catalytic domain of the protein. Similar to *ERBB2* and *MET*, the receptors encoded by these genes transmit signals through the RAS/MEK cascade and when mutated can lead to constitutive activation of oncogenic pathways^{17,19}.

We further examined candidate alterations within the RAS pathway and observed a K57R change in the mitogen-activated protein kinase gene *MAP2K1* in a cetuximab-resistant case. Alterations of *MAP2K1* at the same or nearby residues have been previously described in various cancers, are adjacent to the catalytic domain, and have been shown to confer IL-3-independent cell growth *in vitro*, suggesting that this mutation may be functionally active²⁰. Overall, the enrichment of mutations in these pathways in the resistant tumour grafts was statistically significant ($P < 0.001$, Welch's two-sample *t*-test) and suggests that alterations in any of these members may be sufficient to render cells insensitive to *EGFR* inhibition.

To extend the observations, we analysed 65 cetuximab-naïve samples from patients who were subsequently treated with anti-*EGFR* therapy as part of clinical trials or standard of care. We detected coding alterations in genes known to be involved in *EGFR* therapeutic resistance, including *KRAS*, *NRAS*, *BRAF*, *PIK3CA*, and *PTEN* sequence mutations, and amplification of *MET* and *ERBB2* (a total of 25 cases with mutation in at least one resistance gene). In the remaining 40 cases, we confirmed observations of alterations in several genes with novel resistance mechanisms, including sequence changes in *ERBB2* and *PDGFRA* (Supplementary Tables 1–3).

Although some tumours respond to cetuximab, virtually all patients with CRC develop disease recurrence. In our analyses, 22 tumours were from patients who received cetuximab within 6 months before resection (Supplementary Table 1). We examined whether alterations in these cases may have arisen as acquired (secondary) resistance to therapy. Two of these 22 tumours had somatic sequence changes in *EGFR* (G465R or G465E) affecting domain III of the extracellular portion of the receptor. Structural analyses suggested that these mutations were likely to affect cetuximab binding as they were located at the interface of *EGFR*–cetuximab interaction (Fig. 3a and Extended Data Fig. 2). Interestingly, G465 is structurally adjacent to residue S492 that has been shown, when altered, to interfere with cetuximab binding¹¹ (Fig. 3a). We sequenced pre- and post-therapy specimens for the two patients (CRC104 and CRC177) whose tumours harboured the ectodomain mutations. In both cases, we confirmed the *EGFR* mutations in the post-cetuximab metastases while the original pre-treatment specimens did not have detectable alterations (Fig. 3b, c).

Among patients with CRC with *KRAS* wild-type tumours, only 12–17% have durable responses to anti-*EGFR* monotherapy^{4,6}. We wondered whether such responses may be due to alterations in genes that confer therapeutic sensitivity. *EGFR* was found to be amplified in two tumours that showed either regression (CRC98, 26-fold amplified) or disease stabilization (CRC400, 3-fold amplified) (Fig. 2), consistent with previous observations^{21,22}. Given the importance of *EGFR* signalling in CRC, we analysed other pathway members that were preferentially mutated in responsive tumours and identified *IRS2*, a cytoplasmic adaptor that mediates signalling between receptor

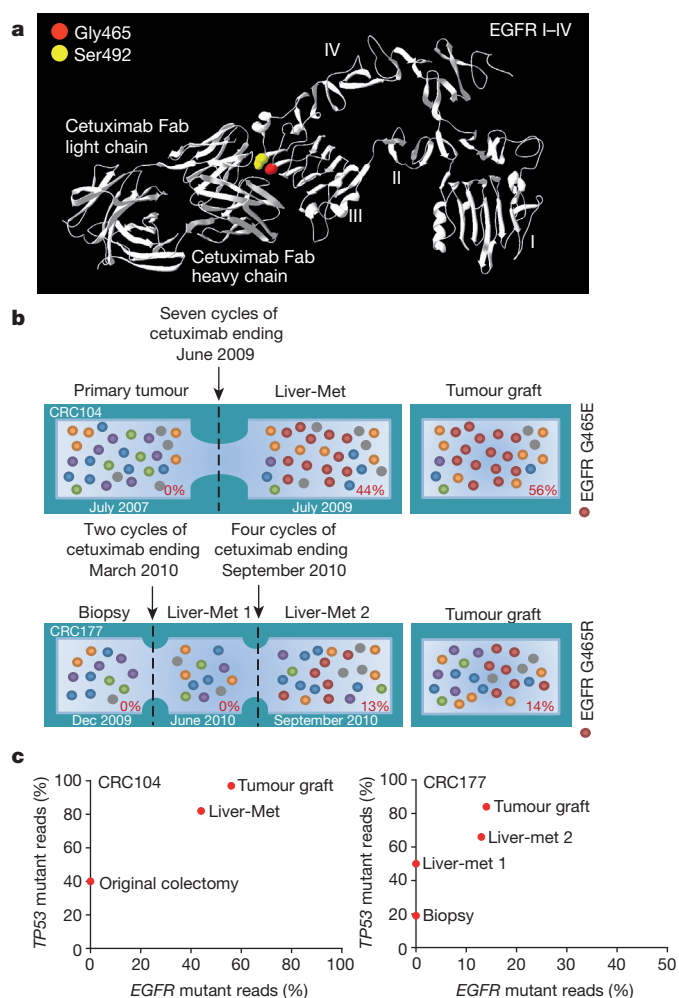


Figure 3 | Genetic alterations involved in secondary resistance to anti-*EGFR* therapy. **a**, The locations of mutations in *EGFR* ectodomain are shown including G465 (red) and the S492 residue known to confer cetuximab resistance¹¹ (yellow). **b**, Evolution of *EGFR* mutations in two CRCs with acquired resistance to cetuximab. Cetuximab-naïve samples were sequenced to investigate the presence of *EGFR* G465 mutations (red) before treatment. For each sample, the fraction of mutant tags is indicated. Met, metastases. **c**, As a control for tumour cellularity, for each lesion the fraction of *TP53* mutant reads (vertical axis) was plotted against the fraction of reads with *EGFR* ectodomain mutations (horizontal axis).

tyrosine kinases and downstream targets (Fig. 2 and Supplementary Table 6) ($P < 0.05$, Welch's two-sample *t*-test). *IRS2* had amplifications or sequence alterations in seven tumours (10%) that showed increased sensitivity or stable disease when treated with cetuximab. Expression analyses of 100 CRC tumour grafts with wild-type *KRAS*, *NRAS*, *BRAF*, and *PIK3CA* identified increased *IRS2* levels as a significant predictor of cetuximab sensitivity (Extended Data Fig. 3). A few tumours that were not responsive to cetuximab harboured *IRS2* alterations together with known resistance changes, including those in *MET* or *BRAF*. These observations suggest that *IRS2* mutations may predict anti-*EGFR* sensitivity in cases without other mechanisms of resistance to *EGFR* therapy. We and others have previously identified alterations in *IRS2* in CRCs and other tumour types, but no reports so far have linked the effects of these alterations to therapeutic sensitivity^{14,23}.

To evaluate the role of these novel alterations, we performed functional assays in NCI-H508, a cetuximab-sensitive CRC cell line that does not harbour known resistance-conferring mutations^{24,25} and displays a threefold gene copy number gain of the *IRS2* gene (Supplementary Tables 3 and 4). We found that ectopic introduction of either *EGFR* G465E or *MAP2K1* K57N into NCI-H508 cells induced resistance to

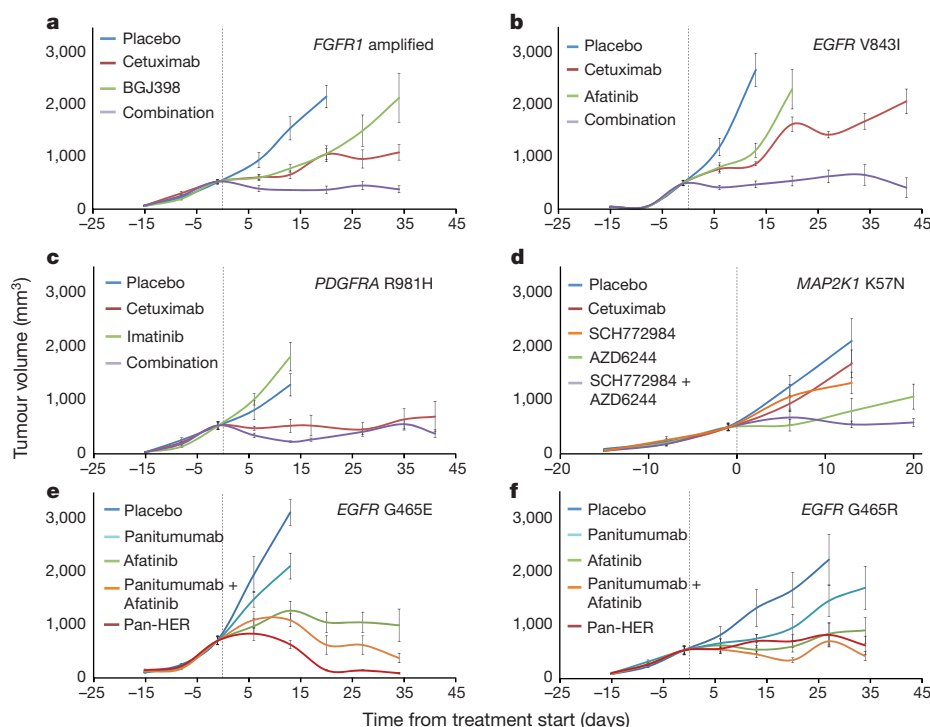


Figure 4 | Therapeutic intervention in preclinical trials to overcome resistance to anti-EGFR antibody blockade. **a–f**, Tumour growth curves in tumour graft cohorts from individual patients with *FGFR1* amplification (CRC477) (**a**), *EGFR* kinase mutation (CRC334) (**b**), *PDGFRA* R981H mutation (CRC525) (**c**), *MAP2K1* K57N mutation (CRC343) (**d**), and *EGFR* ectodomain mutations (**e**, CRC104; **f**, CRC177) treated with placebo or targeted treatments. Mean tumour volumes \pm s.e.m. are shown ($n = 5$ mice per

group for CRC525 and CRC177; $n = 6$ mice per group for all other models). **a, b**, Combination versus cetuximab, $P < 0.01$; **c**, combination versus cetuximab, not significant; **d**, SCH772984 + AZD6244 versus either monotherapy, $P < 0.01$; **e, f**, afatinib, Pan-HER or panitumumab + afatinib versus panitumumab, $P < 0.01$. Statistical analysis was performed by two-way analysis of variance (ANOVA).

EGFR inhibition and increased activation of downstream signals, which were not affected by EGFR blockade (Extended Data Fig. 4a, b). Conversely, knockdown of IRS2 by short hairpin RNA (shRNA) resulted in reduced sensitivity to cetuximab and less pronounced activation of ERK and AKT following ligand stimulation (Extended Data Fig. 4c). This is consistent with the role of IRS2 as a scaffold/adaptor protein that amplifies signals downstream from tyrosine kinase receptors.

Given the poor outcome of patients diagnosed with late-stage CRC, we investigated whether mutant genes observed in individual cases may be clinically actionable using existing or investigational therapies. We identified somatic alterations with potentially actionable consequences in 100 of the 129 patients (77%) (Supplementary Table 8). To test whether any of the identified alterations could be successfully targeted in tumours with cetuximab resistance, we used the tumour grafts to perform proof-of-principle trials for targeted therapies and evaluated the signalling consequences of these therapies *in vivo* (Fig. 4 and Extended Data Figs 5–10). We chose a cetuximab-resistant tumour with *FGFR1* amplification (CRC477) and examined whether inhibition of both *FGFR1* and *EGFR* would be more effective than inhibition of *EGFR* alone. We confirmed resistance to cetuximab alone and, as may be expected using a single inhibitor, the tumour graft was also resistant to monotherapy with the selective *FGFR* kinase inhibitor BGJ398, which is currently in clinical trials (Fig. 4a). However, combination of BGJ398 with cetuximab led to a substantial and durable suppression of tumour growth in all treated mice. This model confirmed that combinatorial therapies may be effective in overcoming EGFR therapeutic resistance in tumours with alterations in other cell-surface receptors.

An analogous approach was used to evaluate the EGFR small-molecule inhibitor afatinib in tumour CRC334 containing sequence change V843I in the protein kinase domain of EGFR. Similar to our observations for *FGFR1* targeting, treating tumour grafts with afatinib or cetuximab alone was not effective but the combination resulted in

marked and long-lasting inhibition of tumour growth (Fig. 4b). We also found that combinations of MEK and ERK inhibitors in tumour graft CRC343 (*MAP2K1* K57N), and the *PDGFR* inhibitor imatinib and cetuximab in tumour graft CRC525 (*PDGFRA* R981H), exerted strong anti-tumour activities (Fig. 4c, d), although the effect was short-lived in the *PDGFRA* mutant tumour. Targeting of *ERBB2* mutations in cetuximab-resistant CRC tumour grafts has been recently demonstrated using dual HER2-targeted therapy in a separate study²⁶. Consistent with the observed higher efficacy of combination therapies, we found that the impact of therapies on downstream signals was stronger when tumours were targeted by drug combinations compared with single-agent treatments (Extended Data Figs 5–10).

Next, we evaluated alternative therapeutic approaches in tumours with secondary cetuximab-resistant alterations in the EGFR ectodomain. Although previous reports have shown that cetuximab-resistant tumours with S492R alterations in EGFR are sensitive to panitumumab¹¹, tumour grafts with the *EGFR* G465E mutation were poorly sensitive to panitumumab (Fig. 4e). Structural analyses indicate that the S492 residue is in the cetuximab binding site within EGFR domain III, while G465 is located in the centre of the region in which the epitopes of both antibodies overlap²⁷ (Extended Data Fig. 2). This lack of sensitivity was not due to absence of *EGFR* dependence as kinase inhibition of EGFR using afatinib resulted in reduction of tumour growth (Fig. 4e). To explore whether EGFR inhibition by antibodies targeting epitopes far from G465 might overcome resistance, we used Pan-HER (Symphogen), a monoclonal antibody mixture that binds EGFR epitopes different from those recognized by cetuximab and panitumumab²⁸ (Extended Data Fig. 2). Pan-HER displayed strong anti-tumour activity in both CRC104 with the *EGFR* G465E mutation (Fig. 4e) and CRC177 with the *EGFR* G465R mutation (Fig. 4f).

Our genomic analyses have detected essentially all previously known mechanisms of resistance to cetuximab in CRC. The results

identify novel candidate mechanisms of primary and secondary resistance through alterations affecting *EGFR*, its downstream signalling pathway, and other cell-surface receptors (Extended Data Fig. 1). These alterations, together with *KRAS*, constitute over three-quarters of cetuximab-resistant tumours and suggest that the vast majority of mechanisms of primary resistance have now been determined and can be identified before the initiation of anti-EGFR treatment.

Some of the mechanisms of resistance to EGFR therapy provide new avenues for intervention, including amplification of *FGFR1* and mutations of *PDGFR1*, *ERBB2*, and *MAP2K1*. As we have shown, combinations of therapies targeting both the protein products encoded by resistance genes as well as EGFR or other signalling partners are likely to be crucial for inhibiting the multiple genetic components within a tumour. Although combinatorial treatments in tumour grafts often led to arrest of tumour growth rather than regression, disease stabilization is prognostically relevant and is the most common consequence of EGFR-targeted therapies in responsive patients with CRC⁴. The high fraction of tumours with actionable alterations suggests that additional combinatorial therapies may be clinically useful for patients with CRC.

An unexpected finding was the identification of *IRS2* alterations as a novel mechanism of sensitivity to anti-EGFR therapy. Our genetic and functional data suggest that *IRS2* alterations may identify tumours that are dependent on receptor signalling and therefore sensitive to its therapeutic inhibition. Consistent with this prediction are reports that *IRS2* amplification is a significant indicator of response to the IGF1R inhibitor figitumumab in colorectal and lung cancer cell lines²⁹. Given the interaction of *IRS2* with multiple cell-surface receptors, we would predict that combinatorial inhibition of these receptors in tumours with *IRS2* alterations may provide additional avenues of intervention in such patients.

This study highlights information that may be obtained through the integration of large-scale genomic and targeted therapeutic analyses in CRC. It provides an unprecedented view into mechanisms of sensitivity as well as primary and secondary resistance to EGFR blockade. This information gives a framework for analysis of responses to targeted therapies in CRC and suggests interventional clinical trials using combinatorial therapies based on potentially actionable alterations.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 17 November 2014; accepted 22 July 2015.

Published online 30 September 2015.

1. Van Cutsem, E., Cervantes, A., Nordlinger, B. & Arnold, D. on behalf of the ESMO Guidelines Working Group. Metastatic colorectal cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* **25** (Suppl. 3), iii1–iii9 (2014).
2. Diaz, L. A. Jr *et al.* The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. *Nature* **486**, 537–540 (2012).
3. Misale, S. *et al.* Emergence of *KRAS* mutations and acquired resistance to anti-EGFR therapy in colorectal cancer. *Nature* **486**, 532–536 (2012).
4. Amado, R. G. *et al.* Wild-type *KRAS* is required for panitumumab efficacy in patients with metastatic colorectal cancer. *J. Clin. Oncol.* **26**, 1626–1634 (2008).
5. De Roock, W. *et al.* Effects of *KRAS*, *BRAF*, *NRAS*, and *PIK3CA* mutations on the efficacy of cetuximab plus chemotherapy in chemotherapy-refractory metastatic colorectal cancer: a retrospective consortium analysis. *Lancet Oncol.* **11**, 753–762 (2010).
6. Tol, J. *et al.* Markers for EGFR pathway activation as predictor of outcome in metastatic colorectal cancer patients treated with or without cetuximab. *Eur. J. Cancer* **46**, 1997–2009 (2010).
7. Sartore-Bianchi, A. *et al.* *PIK3CA* mutations in colorectal cancer are associated with clinical resistance to EGFR-targeted monoclonal antibodies. *Cancer Res.* **69**, 1851–1857 (2009).
8. Bardelli, A. *et al.* Amplification of the MET receptor drives resistance to anti-EGFR therapies in colorectal cancer. *Cancer Discov.* **3**, 658–673 (2013).
9. Bertotti, A. *et al.* A molecularly annotated platform of patient-derived xenografts (“xenopatients”) identifies HER2 as an effective therapeutic target in cetuximab-resistant colorectal cancer. *Cancer Discov.* **1**, 508–523 (2011).
10. Yonesaka, K. *et al.* Activation of *ERBB2* signaling causes resistance to the EGFR-directed therapeutic antibody cetuximab. *Sci. Transl. Med.* **3**, 99ra86 (2011).

11. Montagut, C. *et al.* Identification of a mutation in the extracellular domain of the epidermal growth factor receptor conferring cetuximab resistance in colorectal cancer. *Nature Med.* **18**, 221–223 (2012).
12. Bettegowda, C. *et al.* Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci. Transl. Med.* **6**, 224ra224 (2014).
13. Diaz, L. A. Jr, Sausen, M., Fisher, G. A. & Velculescu, V. E. Insights into therapeutic resistance from whole-genome analyses of circulating tumor DNA. *Oncotarget* **4**, 1856–1857 (2013).
14. Leary, R. J. *et al.* Integrated analysis of homozygous deletions, focal amplifications, and sequence alterations in breast and colorectal cancers. *Proc. Natl Acad. Sci. USA* **105**, 16224–16229 (2008).
15. Barber, T. D., Vogelstein, B., Kinzler, K. W. & Velculescu, V. E. Somatic mutations of EGFR in colorectal cancers and glioblastomas. *N. Engl. J. Med.* **351**, 2883 (2004).
16. Moroni, M. *et al.* Somatic mutation of EGFR catalytic domain and treatment with gefitinib in colorectal cancer. *Ann. Oncol.* **16**, 1848–1849 (2005).
17. Wesche, J., Haglund, K. & Haugsten, E. M. Fibroblast growth factors and their receptors in cancer. *Biochem. J.* **437**, 199–213 (2011).
18. Heinrich, M. C. *et al.* PDGFRA activating mutations in gastrointestinal stromal tumors. *Science* **299**, 708–710 (2003).
19. Dibb, N. J., Dilworth, S. M. & Mol, C. D. Switching on kinases: oncogenic activation of BRAF and the PDGFR family. *Nature Rev. Cancer* **4**, 718–727 (2004).
20. Marks, J. L. *et al.* Novel MEK1 mutation identified by mutational analysis of epidermal growth factor receptor signaling pathway genes in lung adenocarcinoma. *Cancer Res.* **68**, 5524–5528 (2008).
21. Algars, A., Lintunen, M., Carpen, O., Ristamaki, R. & Sundstrom, J. EGFR gene copy number assessment from areas with highest EGFR expression predicts response to anti-EGFR therapy in colorectal cancer. *Br. J. Cancer* **105**, 255–262 (2011).
22. Moroni, M. *et al.* Gene copy number for epidermal growth factor receptor (EGFR) and clinical response to antiEGFR treatment in colorectal cancer: a cohort study. *Lancet Oncol.* **6**, 279–286 (2005).
23. Parsons, D. W. *et al.* Colorectal cancer: mutations in a signalling pathway. *Nature* **436**, 792 (2005).
24. Misale, S. *et al.* Blockade of EGFR and MEK intercepts heterogeneous mechanisms of acquired resistance to anti-EGFR therapies in colorectal cancer. *Sci. Transl. Med.* **6**, 224ra226 (2014).
25. Zanella, E. R. *et al.* IGF2 is an actionable target that identifies a distinct subpopulation of colorectal cancer patients with marginal response to anti-EGFR therapies. *Sci. Transl. Med.* **7**, 272ra212 (2015).
26. Kavuri, S. M. *et al.* HER2 activating mutations are targets for colorectal cancer treatment. *Cancer Discov.* **5**, 832–841 (2015).
27. Voigt, M. *et al.* Functional dissection of the epidermal growth factor receptor epitopes targeted by panitumumab and cetuximab. *Neoplasia* **14**, 1023–1031 (2012).
28. Koefoed, K. *et al.* Rational identification of an optimal antibody mixture for targeting the epidermal growth factor receptor. *MAbs* **3**, 584–595 (2011).
29. Pavlicek, A. *et al.* Molecular predictors of sensitivity to the insulin-like growth factor 1 receptor inhibitor Figitumumab (CP-751,871). *Mol. Cancer Ther.* **12**, 2929–2939 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank S. Angioli, D. Riley, L. Kann, M. Shukla, and C. L. McCord for their assistance with next-generation sequencing analyses, and F. Galimi and S. M. Leto for their help with Sanger sequencing analyses and functional studies. This work was supported by the John G. Ballenger Trust, FasterCures Research Acceleration Award, the European Community's Seventh Framework Programme, the AIRC Italian Association for Cancer Research (Special Program Molecular Clinical Oncology 5 × 1000, project 9970, and Investigator Grants projects 14205 and 15571), American Association for Cancer Research (AACR) – Fight Colorectal Cancer Career Development Award in memory of Lisa Dubow (project 12-20-16-BERT), the Commonwealth Foundation, Swim Across America, US National Institutes of Health grant CA121113, Fondazione Piemontese per la Ricerca sul Cancro-ONLUS (5 × 1000 Italian Ministry of Health 2011), Oncologia Ca' Granda ONLUS, and the SU2C-DCS International Translational Cancer Research Dream Team Grant (SU2C-AACR-DT1415). We acknowledge Merck for a gift of cetuximab. Stand Up To Cancer is a program of the Entertainment Industry Foundation administered by the American Association for Cancer Research. A.B. and L.T. are members of the EuroPDX Consortium.

Author Contributions A.B. and E.P. conceived the project, designed and performed experiments, interpreted results and co-wrote the manuscript. S.J., V.A., V.A., B.L., M.S., J.P., C.A.H., M.N., K.L., F.S., F.C., G.M., E.R.Z., D.R., N.R., A.M., A.M., G.P., M.S., S.M., and A.C. performed experiments, analysed data, prepared tables, or participated in discussion of the results. M.K. and J.L. contributed reagents. K.K.L. undertook all pathological evaluations. C.T., N.N., R.K., and R.S. performed statistical analyses. A.S.-B., S.S., and L.A.D. provided clinically annotated samples and supervised experimental designs. L.T. and V.E.V. conceived the project, supervised experimental designs, interpreted results, and co-wrote the manuscript.

Author Information Sequence data have been deposited at the European Genome-phenome Archive, which is hosted at the European Bioinformatics Institute, under study accession EGAS00001001305. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to V.E.V. (velculescu@jhmi.edu), L.T. (livio.trusolino@ircc.it) or A.B. (andrea.bertotti@ircc.it).

METHODS

Specimens obtained for sequencing analysis. The study population consisted of matched tumour and normal samples from 137 patients with CRC who underwent surgical resection of liver metastases at the Candiolo Cancer Institute, the Mauriziano Umberto I Hospital and the San Giovanni Battista Hospital (all Turin) from 2008 to 2012. Informed consent for research use was obtained from all patients at the enrolling institution before tissue banking, and study approval was obtained from the ethics committees of the three centres. Tumours with *KRAS* alterations at codons 12, 13, and 61 that were detected using Sanger sequencing were not included in the study. From the resected tumour samples, tumour graft models were established as described below. Following exome sequence analyses, eight tumour grafts were detected to have *KRAS* mutations (patients CRC18, CRC58, CRC68, CRC237, CRC312, CRC328, CRC344, CRC382) and were excluded from further analyses. To assess genomic similarity between tumour grafts and the tumours from which they were derived, 18 pre-implantation liver metastases were analysed through targeted next-generation sequencing and compared with the corresponding tumour grafts. Pathological analyses showed that tumour cellularity of the metastatic samples ranged from 15% to 90% (average 59%), supporting the need for enrichment of tumour cells through growth of tumour grafts. Targeted next-generation sequencing revealed that all the clonal alterations identified in these tumour grafts were present in the tumours from which they were derived (Supplementary Table 3), similar to previous comparisons of tumour grafts and primary tumours in CRC³⁰. To extend observations of alterations in resistance mechanisms that we had identified in tumour grafts, an additional 65 patient-derived cetuximab-naïve clinical samples from patients who were subsequently treated with EGFR blockade through standard of care or various clinical trials, including NCT00113763, NCT00891930, NCT00113776, and NCT01126450 (ref. 31), were analysed through targeted genomic analyses (Supplementary Table 9). Available clinical information for all samples is shown in Supplementary Table 2.

Tumour graft models and *in vivo* treatments. Tissue from hepatic metastasectomy in affected individuals was fragmented and either frozen or prepared for implantation as described previously^{9,32}. Non-obese diabetic/severe combined immunodeficient (NOD/SCID) female mice (4–6 weeks old) were used for tumour implantation. Nucleic acids were isolated from early-passaged tumour grafts. The remaining tumour graft material was further passaged and expanded into treatment groups. The size of the animal groups ($n = 5$ or 6) and schedule of measurements (one measurement at baseline and three to five sequential weekly measurements on treatment) were calculated to detect a difference of tumour volumes between mice treated with monotherapy and mice treated with combination therapies. Therefore, three comparisons were considered as primary objective for each experiment. To preserve a family-wise error of 5% (one side), a Bonferroni correction was applied and a type 1 error of 0.017 for each of the three comparisons was considered. This resulted in a power of 80% to detect a standardized comparison of 0.70. Animals with established tumours defined as an average volume of 400 mm³ were randomized and treated with vehicle or drug regimens, either as a single-agent or in combination as indicated: cetuximab (Merck), 20 mg/kg twice a week intraperitoneally; BGI398 (Sequoia Research Products), 30 mg/kg once-daily by oral gavage; imatinib (Sequoia Research Products), 100 mg/kg once-daily by oral gavage; panitumumab (Amgen), 20 mg/kg twice a week intraperitoneally; afatinib (Sequoia Research Products), 20 mg/kg once-daily by oral gavage; AZD6244 (Sequoia Research Products), 25 mg/kg once-daily by oral gavage; SCH772984 (ChemieTek), 75 mg/kg/once daily intraperitoneally; Pan-HER (SympHogen), 60 mg/kg twice a week intraperitoneally. To evaluate sensitivity to cetuximab monotherapy, each tumour graft was evaluated at 3 and 6 weeks in 12 or 24 mice (depending on individual models) that were randomized to treatment and control arms at a 1:1 ratio. For assessment of tumour response to therapy, we used volume measurements normalized to the tumour graft volume at the time of initiation of cetuximab treatment. Tumour grafts were classified as follows: (1) tumour regression with a decrease of at least 35% in tumour volume (39 cases, 34%); (2) disease progression with at least a 35% increase in tumour volume (36 cases, 31%); and (3) disease stabilization with a tumour graft volume at levels < 35% growth and < 35% regression (41 cases, 35%). Tumours displaying regression or stabilization continued treatment for additional 3 weeks. Tumour size was evaluated once a week by calliper measurements and the approximate volume of the mass was calculated. Statistical significance for tumour volume changes was calculated using mixed-design ANOVA (repeated measures) when all mice were available for measurements in each treatment group at each time point, and two-way ANOVA when one or more mice died accidentally over the course of the experiments. Results were considered interpretable when at least half of mice per treatment group ($n = 3$) survived until the pre-specified endpoints (minimum, 3 weeks of treatment). All mice alive at the endpoint were included in the analysis (CRC477: six mice treated with placebo or cetuximab, four mice

treated with BGI398, three mice treated with cetuximab + BGI398; CRC334: five mice treated with cetuximab + afatinib, six mice per treatment group in all other arms; CRC525: five mice per treatment group in all arms; CRC343: five mice treated with AZD6244 + SCH772984, six mice per treatment group in all other arms; CRC104: four mice treated with panitumumab + afatinib, six mice per treatment group in all other arms; CRC177: five mice per treatment group in all arms). Operators allocated mice to the different treatment groups during randomization but were blinded during measurements. *In vivo* procedures and related biobanking data were managed using the Laboratory Assistant Suite (LAS), a web-based proprietary data management system for automated data tracking³³. All experiments were conducted with approval from the Animal Care Committee of the Candiolo Cancer Institute, in accordance with Italian legislation on animal experimentation.

Sample preparation and next-generation sequencing. DNA was extracted from patients' tumours, early-passage tumour grafts developed from liver metastases, normal samples (adjacent non-cancerous liver or peripheral blood), and from normal tissue of the same mouse strain as those used to grow the xenografts using the Qiagen DNA FFPE tissue kit or Qiagen DNA blood mini kit. Additional analyses were performed for CRC334 after afatinib anti-EGFR therapy and tumour graft regrowth (indicated in footnote of Supplementary Table 4). Genomic DNA from tumour and normal samples were fragmented and used for Illumina TruSeq library construction (Illumina) according to the manufacturer's instructions or as previously described³⁴. Exonic or targeted regions were captured in solution using the Agilent SureSelect version 4 kit or a custom targeted panel according to the manufacturer's instructions (Agilent) (Supplementary Table 9). The captured library was then purified with a Qiagen MinElute column purification kit and eluted in 17 µl of 70 °C EB to obtain 15 µl of captured DNA library. The captured DNA library was amplified in the following way: eight 30 µl PCR reactions each containing 19 µl of H₂O, 6 µl of 5 × Phusion HF buffer, 0.6 µl of 10 mM dNTP, 1.5 µl of DMSO, 0.30 µl of Illumina PE primer 1, 0.30 µl of Illumina PE primer 2, 0.30 µl of Hotstart Phusion polymerase, and 2 µl of captured exome library were set up. The PCR program used was as follows: 98 °C for 30 s; 14 cycles (exome) or 16 cycles (targeted) of 98 °C for 10 s, 65 °C for 30 s, 72 °C for 30 s; and 72 °C for 5 min. To purify PCR products, a NucleoSpin Extract II purification kit (Macherey-Nagel) was used following the manufacturer's instructions. Paired-end sequencing, resulting in 100 bases from each end of the fragments for exome libraries and 100 or 150 bases from each end of the fragment for targeted libraries, was performed using Illumina HiSeq 2000/2500 and Illumina MiSeq instrumentation (Illumina).

Primary processing of next-generation sequencing data and identification of putative somatic mutations. Somatic mutations were identified using VariantDx³⁴ custom software for identifying mutations in matched tumour and normal samples. Before mutation calling, primary processing of sequence data both for tumour and for normal samples was performed using Illumina CASAVA software (version 1.8), including masking of adaptor sequences. Sequence reads were aligned against the human reference genome (version hg18) using ELAND with additional realignment of select regions using the Needleman-Wunsch method³⁵. Candidate somatic mutations, consisting of point mutations, insertions, and deletions, were then identified using VariantDx across the either the whole exome or regions of interest. VariantDx examines sequence alignments of tumour samples against a matched normal while applying filters to exclude alignment and sequencing artefacts. In brief, an alignment filter was applied to exclude quality failed reads, unpaired reads, and poorly mapped reads in the tumour. A base quality filter was applied to limit inclusion of bases with reported Phred quality score > 30 for the tumour and > 20 for the normal. A mutation in the tumour was identified as a candidate somatic mutation only when (1) distinct paired reads contained the mutation in the tumour, (2) the number of distinct paired reads containing a particular mutation in the tumour was at least 2% of the total distinct read pairs for targeted analyses and 10% of read pairs for exome, (3) the mismatched base was not present in > 1% of the reads in the matched normal sample as well as not present in a custom database of common germline variants derived from dbSNP, and (4) the position was covered in both the tumour and normal. Mutations arising from misplaced genome alignments, including paralogous sequences, were identified and excluded by searching the reference genome. Potential alterations were compared with mouse sequences from experimentally obtained mouse whole-exome and targeted sequence data as well as the reference mouse genome (mm9) to remove mouse-specific variants. Candidate somatic mutations were further filtered on the basis of gene annotation to identify those occurring in protein coding regions. Functional consequences were predicted using snpEff and a custom database of CCDS, RefSeq and Ensembl annotations using the latest transcript versions available on hg18 from UCSC (<https://genome.ucsc.edu/>). Predictions were ordered to prefer transcripts with canonical start and stop codons and CCDS or RefSeq transcripts over Ensembl when available.

Finally, mutations were filtered to exclude intronic and silent changes, while retaining mutations resulting in missense mutations, nonsense mutations, frame-shifts, or splice-site alterations. A manual visual inspection step was used to further remove artefactual changes. Amplification analyses were performed using the digital karyotyping approach³⁶ by comparing the number of reads mapping to a particular gene with the average number of reads mapping to each gene in the panel, along with a minor allele fraction analysis of heterozygous single nucleotide polymorphisms contained within each gene. For comparison of somatic alterations in tumour graft and pre-implantation material, we considered all alterations where the mutation was present in at least 20% of the read pairs in the tumour graft samples. To evaluate whether mutant genes observed in individual cases could be clinically actionable using existing or investigational therapies, we examined altered genes that were associated with (1) US Food and Drug Administration-approved therapies for oncological indications, (2) therapies in published prospective or retrospective clinical studies, and (3) ongoing clinical trials for patients with CRC or other tumour types.

Gene expression analyses. Data were obtained using a HumanHT-12 version 4 Illumina beadarray technology. Following data normalization, genes were collapsed to the probe displaying highest mean signal. Gene expression values were then log₂-transformed and centred to the median (Supplementary Table 10). IRS2 expression in 100 tumour grafts with wild-type forms of *KRAS*, *NRAS*, *BRAF*, and *PIK3CA* was compared with cetuximab response by one-way ANOVA and Bonferroni's multiple comparisons test.

Statistical analyses for genes with somatic alterations. Using the approach previously described³⁷, we analysed 24,334 somatic mutations (non-synonymous and synonymous single-base substitutions plus indels) identified in the protein coding sequence of 127 tumour graft samples, after samples with *KRAS* hotspot mutations (codons 12 or 13) and those with a mutator phenotype were excluded. We implemented the following statistical framework to identify significantly mutated genes by incorporating background mutation rates, gene length, and base composition.

Inspired by previous works^{38,39}, our model defines gene-specific background mutation rates, which capture exome-wide as well as gene-specific sequence-based parameters. We define eight exhaustive and disjoint sequence-based dinucleotide contexts: C in CpG, G in CpG, C in TpC, G in GpA, and all other A, G, C, T. We represent the occurrences of each context in the entire protein coding sequence by N_i , and in each gene of interest by g_i . Subsequently, we identify the dinucleotide context for all single-base substitution somatic mutations identified and derive the counts of mutations in each context over all CDS (protein coding sequence) (n_i). We derive the expected probability of observing a mutation in a base occurring in the CDS of a gene of interest as follows:

$$P_{\text{mut}} = \frac{\sum_{i=1}^I g_i f_i}{\sum_{i=1}^I g_i} \quad (1)$$

$$f_i = \frac{n_i}{N_i} \quad (2)$$

where f_i denotes the fraction of bases in dinucleotide context i in the entire CDS, where a mutation has been observed. The context parameters N_i and g_i are defined as the total number of occurrences of each context sequenced across all of the samples; therefore following the simplifying assumption of full coverage of the entire protein coding sequence, and assuming K samples total, these parameters will be K times those of a single haploid exome.

Following the definition of f_i , we derive the background probability of observing at least $m_{g,\text{obs}}$ mutations in a gene of interest, using the binomial tail probability of L_g trials with $m_{g,\text{obs}}$ successes and P_{mut} probability of success in each trial. Here, L_g represents the length of the CDS of the gene times the number of samples.

$$P_{\text{req}}^{\text{mut}} = P(m_{g,\text{mut}} \geq m_{g,\text{obs}}) = \sum_{j=m_{g,\text{obs}}}^{L_g} \binom{L_g}{j} P_{\text{mut}}^j (1 - P_{\text{mut}})^{L_g-j} \quad (3)$$

We use an equivalent formulation to model the statistical significance of observing $q_{g,\text{obs}}$ insertions/deletions (indels) in a gene of interest. The background indel frequency (P_{indel}) is defined as the number of indels recovered in the entire CDS of the sequenced samples divided by the length of the entire CDS available in these samples.

$$P_{\text{req}}^{\text{indel}} = P(q_{g,\text{indel}} \geq q_{g,\text{obs}}) = \sum_{j=q_{g,\text{obs}}}^{L_g} \binom{L_g}{j} P_{\text{indel}}^j (1 - P_{\text{indel}})^{L_g-j} \quad (4)$$

The two statistical tests described above (equations (3) and (4)) reflect the significance of mutation counts in a gene, but are blind to the protein-level consequence of mutations. To capture the impact of mutation on protein, we apply an extension of the tests above that examines enrichment for non-synonymous mutations in the

set of single-base substitution mutations identified in a gene of interest. We define a background, gene-specific ratio of non-synonymous to synonymous (NS/S) mutations, given the exome-wide NS/S ratio in each dinucleotide context (r_i) and the sequence composition of each gene as follows. Note that g_i has the same definition as in equation (1).

$$r_g = \frac{\sum_{i=1}^I r_i g_i}{\sum_{i=1}^I g_i} \quad (5)$$

Given the NS/S ratio for a gene of interest, the probability of an observed mutation in the gene being non-synonymous is

$$P_{g,\text{NS}} = \frac{r_g}{(r_g + 1)} \quad (6)$$

Following this step, the binomial tail probability of observing $m_{g,\text{obs}}^{\text{NS}}$ from the total of $m_{g,\text{obs}}$ mutations in a gene of interest is:

$$P_{\text{composition}}^{\text{mut}} = P(m_{g,\text{mut}}^{\text{NS}} \geq m_{g,\text{obs}}^{\text{NS}}) = \sum_{j=m_{g,\text{obs}}^{\text{NS}}}^{m_{g,\text{obs}}} \binom{m_{g,\text{obs}}}{j} P_{g,\text{NS}}^j (1 - P_{g,\text{NS}})^{m_{g,\text{obs}}-j} \quad (7)$$

The three test statistics (equations (3), (4) and (7)) rely on three distinct measures for calling a gene significantly mutated: the counts of single-base substitutions, the counts of indels, and the relative counts of non-synonymous to synonymous single-base substitutions. Assuming the independence of these measures, given gene-specific parameters of g_i and L_g , we combine them using Fisher's combined probability test to derive a measure of overall significance for each gene of interest (combined P value). We acknowledge the fact that Fisher's combined probability test is best suited to P values derived from continuous probability distribution functions; however, it has been shown that its application to P values derived from discrete probability distributions results in conservative estimates of P value.

Finally, we apply Bonferroni and Benjamini-Hochberg's correction method to combined P values to control for multiple testing.

Statistical analyses for therapeutic resistance or sensitivity. Statistical models for tumour growth were implemented for each of four mutation profiles that were correlated to resistance or sensitivity to cetuximab treatment. Group A samples had *ERBB2* mutations and/or amplification, *MET* amplification, *EGFR* mutations affecting the ectodomain or kinase domain, *NRAS* mutation, *BRAF* V600E, *FGFR1* amplification, *PDGFRA* mutations affecting the kinase domain and *MAP2K1* K57N. Group B samples had *ERBB2* mutations, *EGFR* mutations affecting the ectodomain or kinase domain, *FGFR1* amplification, *PDGFRA* mutations affecting the kinase domain or *MAP2K1* K57N. Group C samples had amplification of *EGFR* or a mutation or amplification of *IRS2*, while group D samples had amplification or mutation of *IRS2*. As *IRS2* alterations are likely to be predictive of anti-EGFR response in cases without other mechanisms of resistance to EGFR therapy, we excluded two samples that harboured a *MET* amplification or *BRAF* mutation from groups C and D. For each group, Wilcoxon rank sum and Welch's two-sample t -tests were used to evaluate differences in the mean tumour growth between samples with mutation and those without.

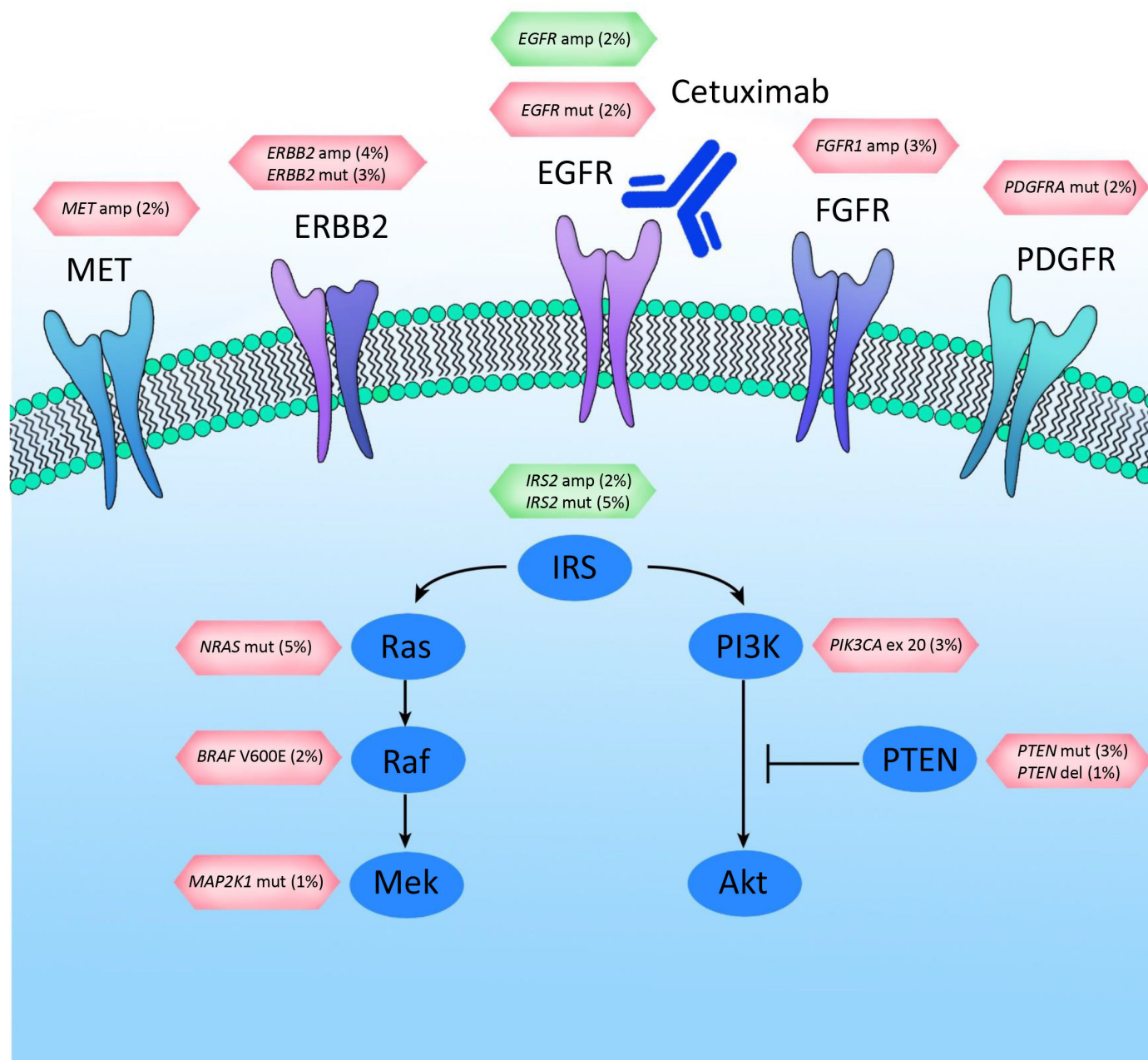
Protein structure modelling. The crystal structure of the extracellular domain of the epidermal growth factor receptor in complex with the Fab fragment of cetuximab was retrieved from the Protein Data Bank (accession number 1YY9). This Protein Data Bank entry contains a complex of three biomacromolecules including the extracellular portion of EGFR, cetuximab Fab Light chain, and cetuximab Fab Heavy chain. The EGFR-cetuximab complex was visualized using Deep View Swiss-pdbviewer (SPDBV_4.10_PC).

Cell cultures, plasmids, antibodies, and biological assays. NCI-H508 and 293T cells were obtained from ATCC and cultured in RPMI 1640 and Iscove medium, respectively. Cell lines were authenticated for genetic identity by short tandem repeat profiling (Cell ID, Promega) and routinely PCR-tested for mycoplasma contamination (Minerva Biolabs). EGFR G465E and MAP2K1 K57N in the PS100069 lentiviral vectors were custom-cloned by and purchased from OriGene. The MISSION lentiviral pLKO.1-puro shRNA vector targeting *IRS2* (target sequence: GTGAAGATCTGTCTGGCTTTA), as well as the non-targeting control vector, were purchased from Sigma. All vectors were produced by lipofectAMINE 2000 (Life Technologies)-mediated transfection of 293T cells. Primary antibodies included rabbit anti-phospho-Tyr1068-EGFR (ab5644) (Abcam); rabbit anti-EGFR (D38B1), rabbit anti-IRS2 (L1326), rabbit anti-phospho-Ser473-AKT (D9E), rabbit anti-AKT (11E7), rabbit anti-phospho-Thr202/Tyr204-ERK (D13.14.4E), rabbit anti-ERK (137F5) (Cell Signaling Technology); mouse anti-DDK (4C5) (Origene); and mouse anti-tubulin (DM1A) (Sigma-Aldrich). Proliferative response was assessed with an ATP content assay as an indicator of cellular viability. On day 0, cells were plated at clonal density (20 cells per microlitre) in complete medium. On day 1, serially diluted cetuximab or

vehicle (PBS) was added to the cells. On day 6, cell viability was measured by CellTiter-Glo (Promega) using Victor X4 (PerkinEmler) or GloMax (Promega) microplate luminometers.

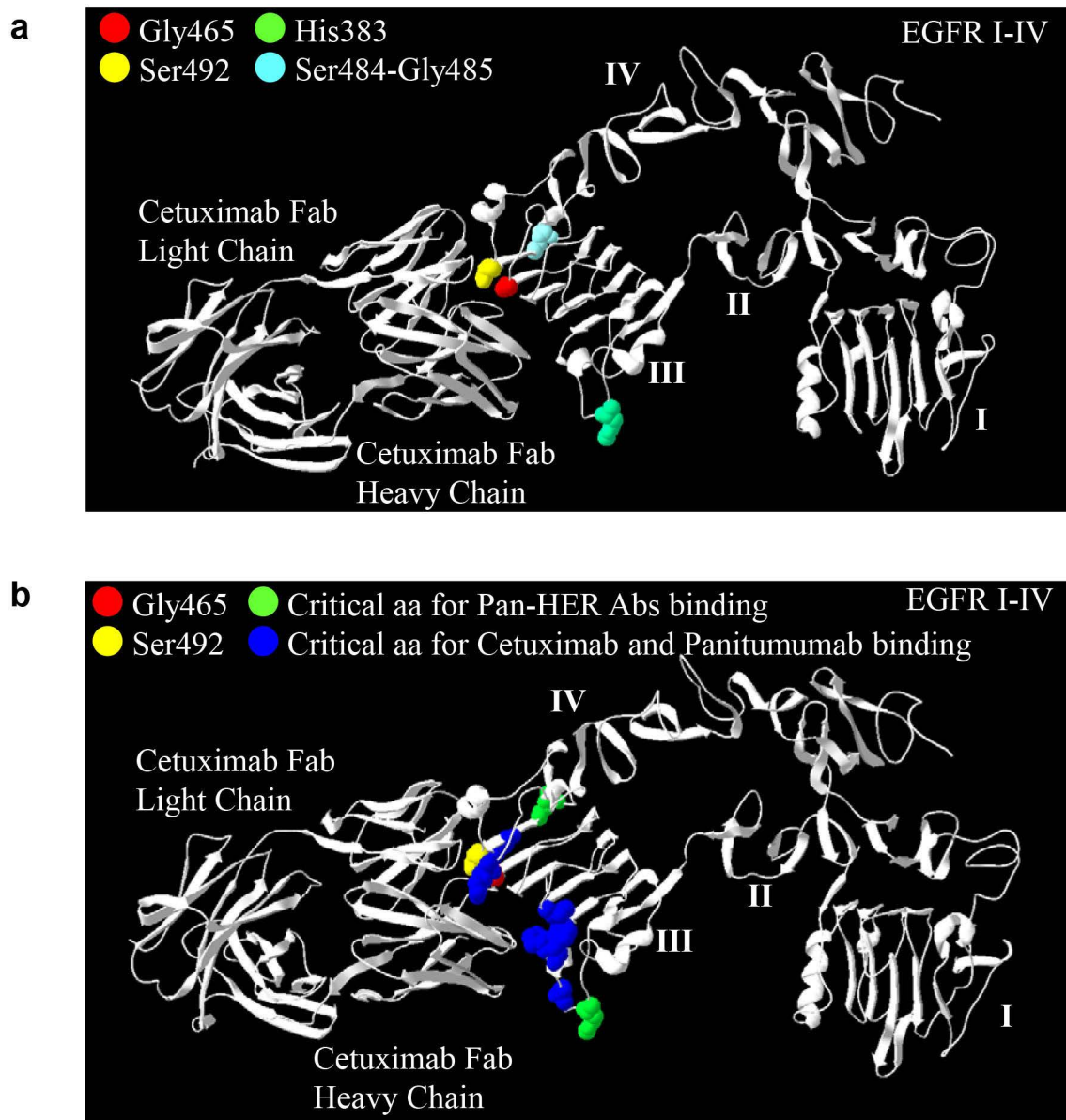
Pharmacodynamic analyses. Tumour grafts were embedded in paraffin and subjected to immunoperoxidase staining with rabbit monoclonal antibodies against phospho-S6 (Ser235/236, clone D57.2.2E, Cell Signaling Technology) or phospho-ERK1/2 (Thr202/Tyr204, clone D13.14.4E, Cell Signaling Technology). After incubation with secondary antibodies, immunoreactivities were revealed by incubation in DAB chromogen (Dako). Images were captured with the Leica LAS EZ software using a Leica DM LB microscope. For morphometric quantitation, five fields per section at $\times 40$ magnification from two tumours from two different mice for each treatment modality ($n = 10$) were analysed using ImageJ. Immunoreactivity for phospho-ERK and phospho-S6 was quantified by spectral segmentation of images in two layers: one layer excluded stroma and empty spaces (such as lumens); the second layer measured DAB positivity. The percentage of immunoreactive cells was calculated as DAB positivity divided by total cancer cell area. Software outputs were manually verified by visual inspection of digital images.

30. Jones, S. *et al.* Comparative lesion sequencing provides insights into tumor evolution. *Proc. Natl Acad. Sci. USA* **105**, 4283–4288 (2008).
31. Siena, S. *et al.* Phase II open-label study to assess efficacy and safety of lenalidomide in combination with cetuximab in KRAS-mutant metastatic colorectal cancer. *PLoS One* **8**, e62264 (2013).
32. Galimi, F. *et al.* Genetic and expression analysis of MET, MACC1, and HGF in metastatic colorectal cancer: response to met inhibition in patient xenografts and pathologic correlations. *Clin. Cancer Res.* **17**, 3146–3156 (2011).
33. Baralis, E., Bertotti, A., Fiori, A. & Grand, A. LAS: a software platform to support oncological data management. *J. Med. Syst.* **36** (Suppl. 1), S81–S90 (2012).
34. Jones, S. *et al.* Personalized genomic analyses for cancer mutation discovery and interpretation. *Sci. Transl. Med.* **7**, 283ra253 (2015).
35. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).
36. Leary, R. J., Cummins, J., Wang, T. L. & Velculescu, V. E. Digital karyotyping. *Nature Protocols* **2**, 1973–1986 (2007).
37. Jiao, Y. *et al.* Exome sequencing identifies frequent inactivating mutations in BAP1, ARID1A and PBRM1 in intrahepatic cholangiocarcinomas. *Nature Genet.* **45**, 1470–1473 (2013).
38. Sjoblom, T. *et al.* The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268–274 (2006).
39. Kan, Z. *et al.* Diverse somatic mutation patterns and pathway alterations in human cancers. *Nature* **466**, 869–873 (2010).



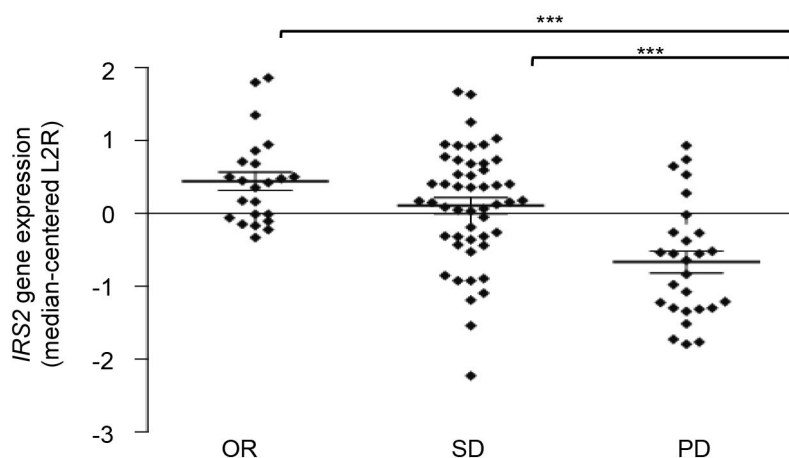
Extended Data Figure 1 | EGFR signalling pathway genes involved in cetuximab resistance or sensitivity. Altered cell-surface receptors or members of RAS or PI3K pathways identified in this study are indicated. Somatic alterations related to resistance or sensitivity are highlighted in red or green boxes, respectively. The percentages indicate the fraction of *KRAS* wild-type

tumours containing the somatic alterations in the specified genes. For the following genes a subset of alterations are indicated: *PDGFRA* kinase domain mutations; *EGFR* ecto- and kinase domain mutations and amplifications.



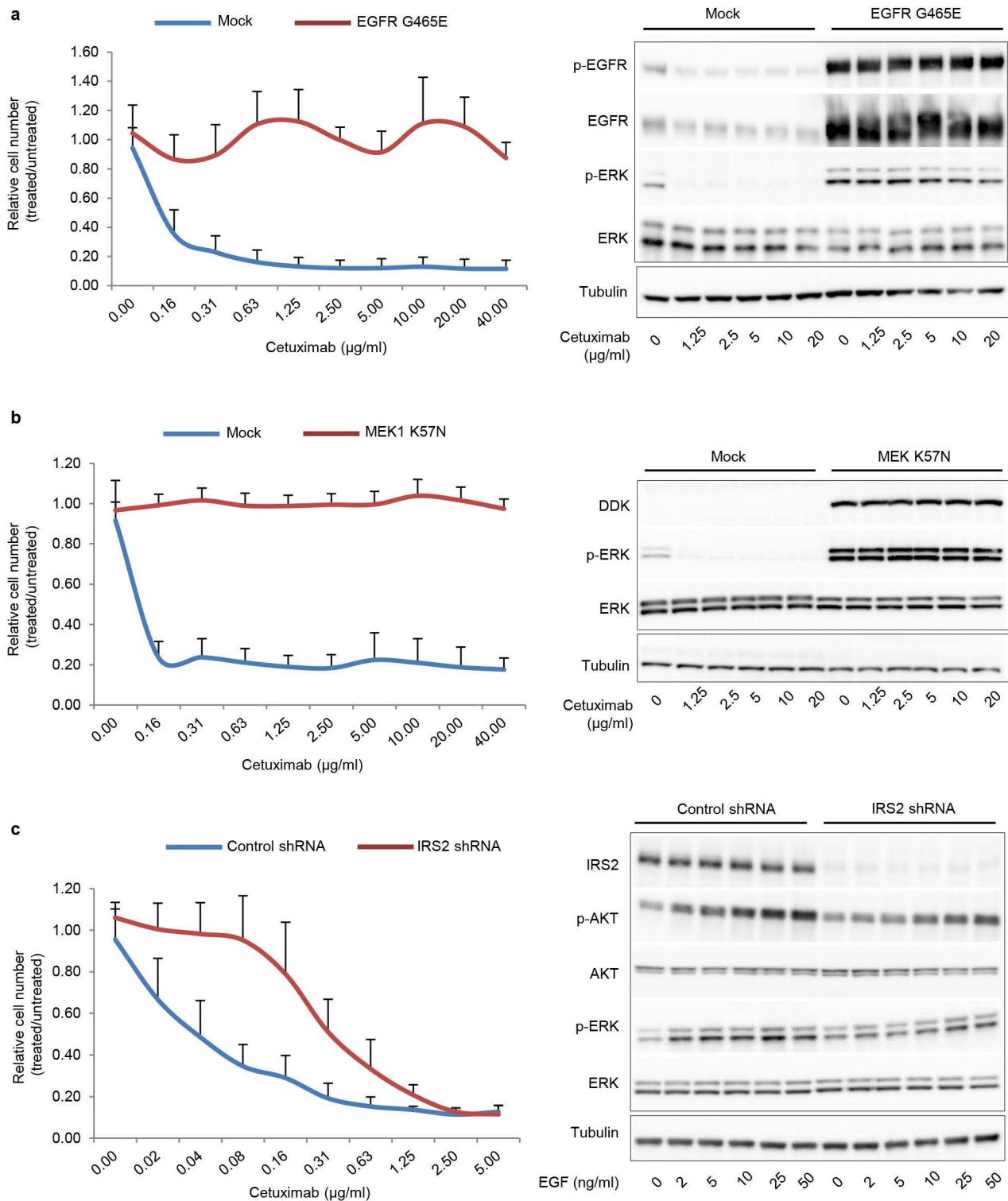
Extended Data Figure 2 | Pan-HER monoclonal antibody mixture binds epitopes different from those recognized by cetuximab. **a**, The H383 (green) and the S484/G485 (light blue) residues in EGFR domain III are critical for the binding of Pan-HER anti-EGFR antibodies 1277 and 1565, respectively²⁸. Antibodies 1277 and 1565 (ref. 28) bind to an epitope distinct from that of cetuximab, which may contribute to the superior tumour growth inhibition in the presence of mutations at residue 465. Mutations identified in this study affecting G465 (red) and the S492 amino acid (yellow) previously reported to confer cetuximab resistance¹¹ are shown for reference. Similarly to mutations

affecting S492, the alterations at 465 that we identified in this study (G465R and G465E) involve changes from a non-polar uncharged side chain to large electrically charged arginine or glutamic acid residues, respectively, and predict resistance to cetuximab. **b**, Critical EGFR amino acids selectively recognized by both cetuximab and panitumumab as determined by phage screening are shown in blue and include P373, K467, P411, K489, D379, F376 (ref. 27). Residue G465 is in close proximity to K467 and other residues that have been shown to influence the binding of both cetuximab and panitumumab²⁷.



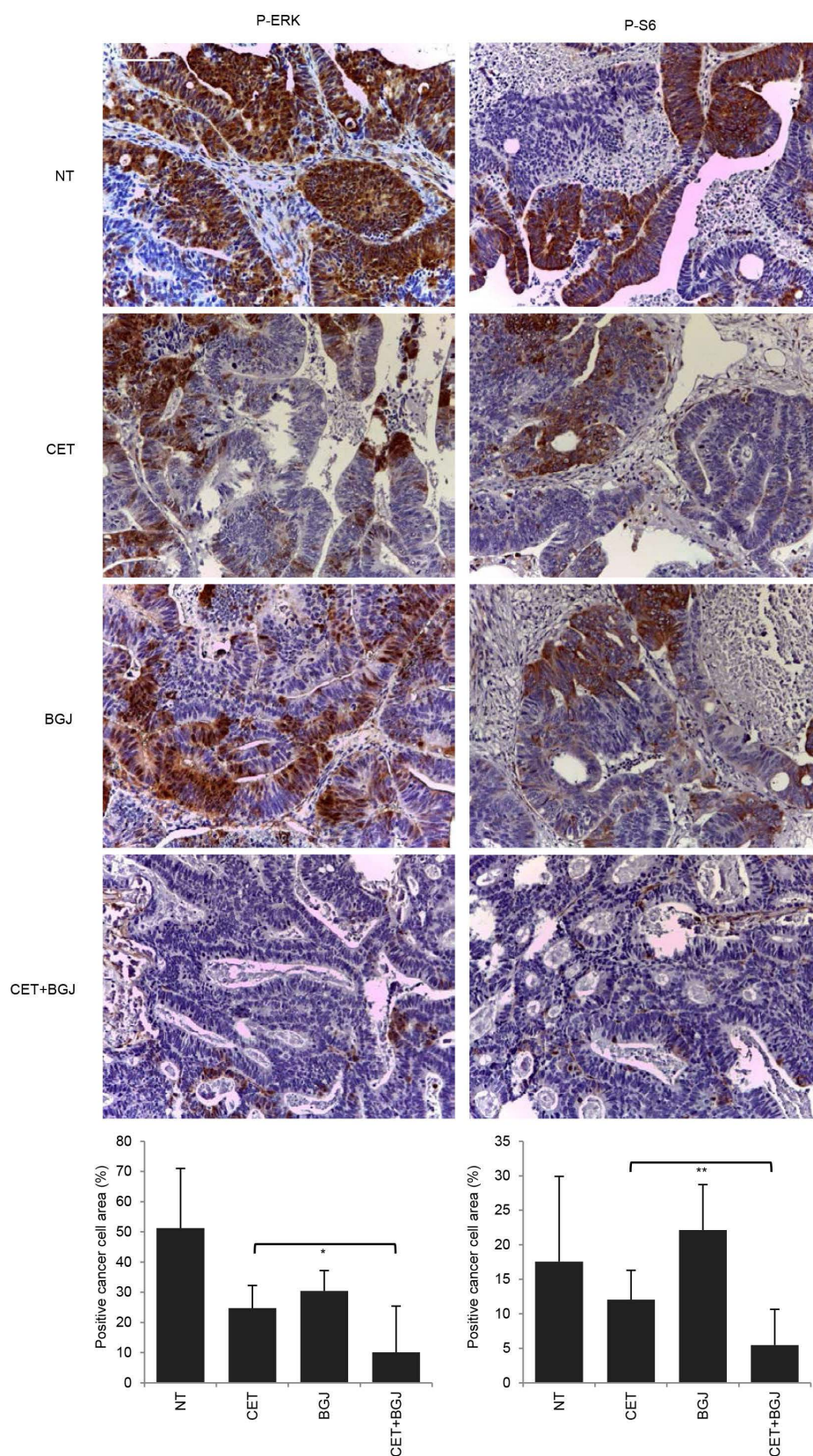
Extended Data Figure 3 | Expression of *IRS2* according to response categories in tumour graft models. Results were obtained using Illumina-based oligonucleotide microarrays in 100 tumour grafts that had no mutations in the *KRAS*, *NRAS*, *BRAF*, or *PIK3CA* genes. Response categories are defined

in the main text. OR, objective response; SD, stable disease; PD, progressive disease. $P < 0.001$ for OR compared with PD and SD compared with PD by one-way ANOVA and Bonferroni's multiple comparison test. *IRS2* expression values are shown in Supplementary Table 10.



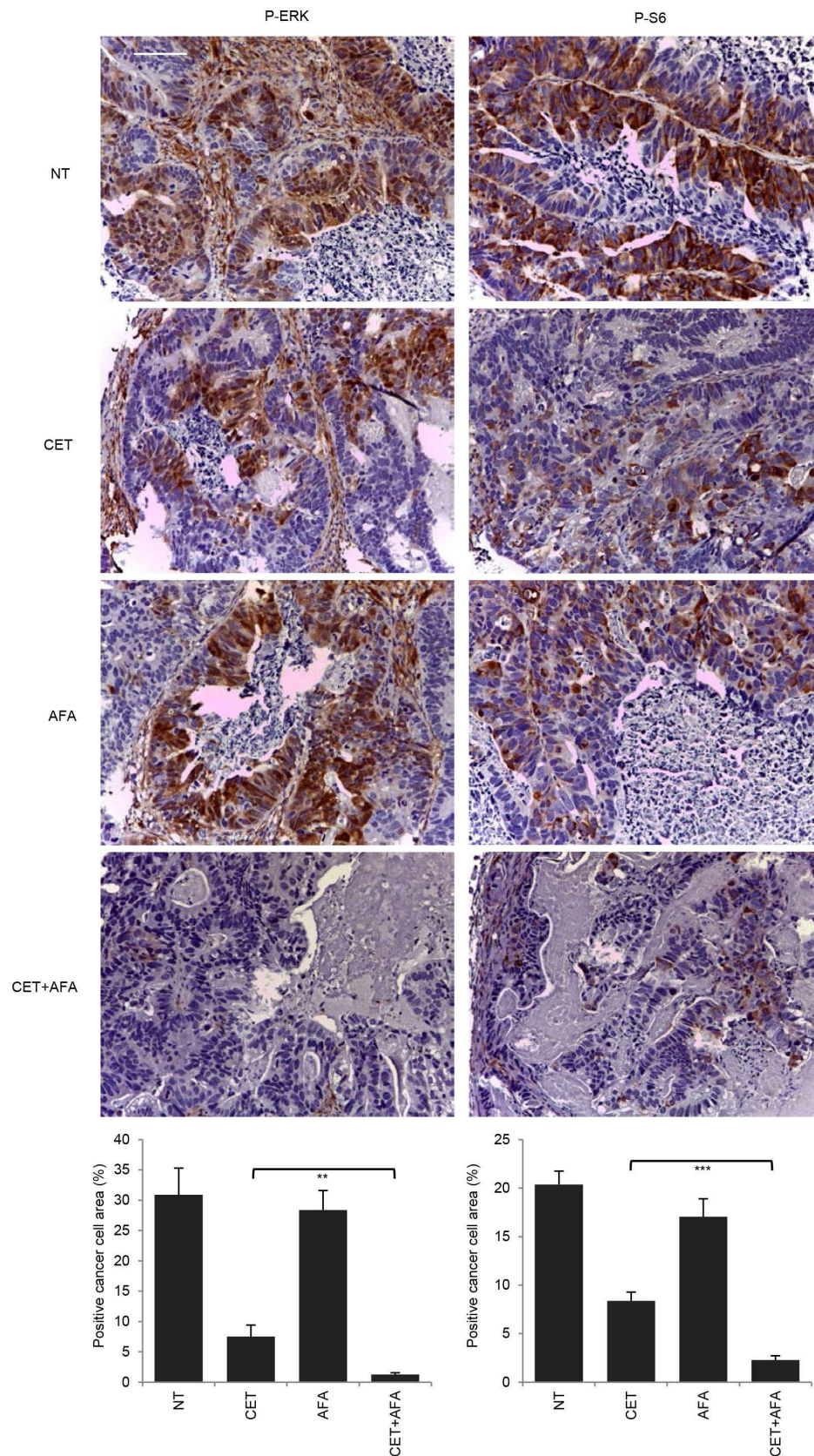
Extended Data Figure 4 | Functional studies of genetic alterations associated with cetuximab response. **a, b,** Ectopic expression of mutations that correlated with resistance to EGFR blockade prevented responsiveness to cetuximab. NCI-H508 cells expressing EGFR G465E (**a**, left) or DDK-tagged MAP2K1 K57N (**b**, left) were refractory to cetuximab in dose-dependent viability assays after 6 days of treatment. Results are the means \pm s.d. of two independent experiments performed in biological triplicates ($n = 6$) for EGFR G465E and three independent experiments performed in biological triplicates ($n = 9$) for MAP2K1 K57N compared with mock vector controls. Biochemical responses of NCI-H508 EGFR G465E (**a**, right) and NCI-H508 MAP2K1 K57N (**b**, right) treated with cetuximab for 24 h were documented by western blot

analyses. **c,** Genetic silencing of IRS2 (IRS2 shRNA) in NCI-H508 cells reduced sensitivity to cetuximab in dose-dependent viability assays (left). Results are the means \pm s.d. of two independent experiments performed in biological triplicates ($n = 6$). In biochemical studies using western blot analyses (right), IRS2 knockdown attenuated EGF-dependent activation of AKT (P-AKT) and ERK (P-ERK). Cells were treated for 10 min with the indicated concentrations of EGF. Tubulin was used as a loading control. Western blots for total EGFR, ERK, and AKT proteins were run with the same lysates as those used for anti-phosphoprotein detection but on different gels. All western blots are representative of two independent experiments.



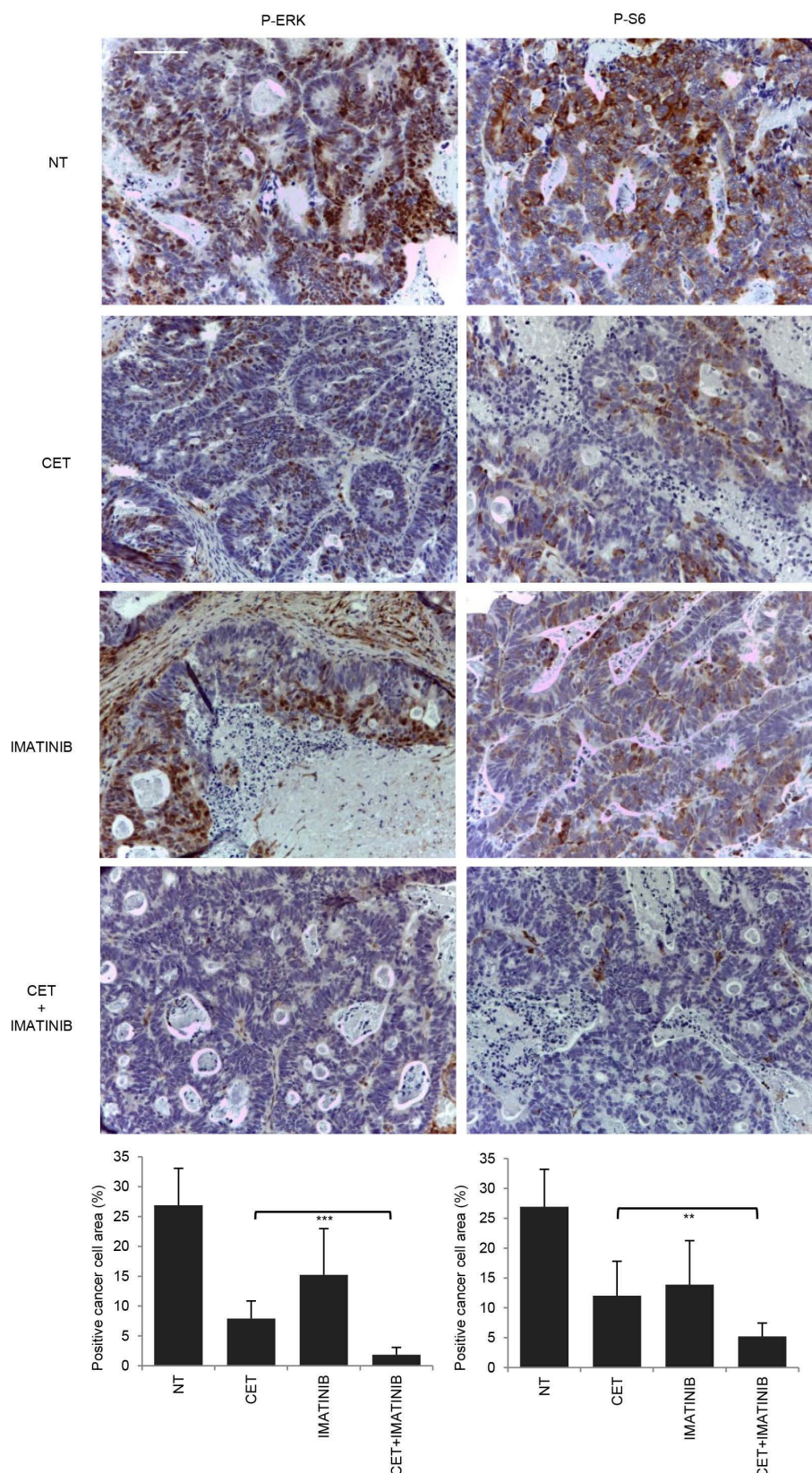
Extended Data Figure 5 | Signalling consequences of FGFR inhibition in FGFR1-amplified CRC477. Immunohistochemistry with the indicated antibodies and morphometric quantitations of representative tumours at the end of treatment. Results are the means \pm s.d. of five fields ($\times 40$) from two

tumours for each experimental point ($n = 10$). Scale bar, 300 μ m. P-ERK, phospho-ERK; P-S6, phospho-S6. NT, not treated (vehicle); CET, cetuximab; BGJ, BGJ398. * $P < 0.05$; ** $P < 0.01$ by two-tailed Student's t -test.



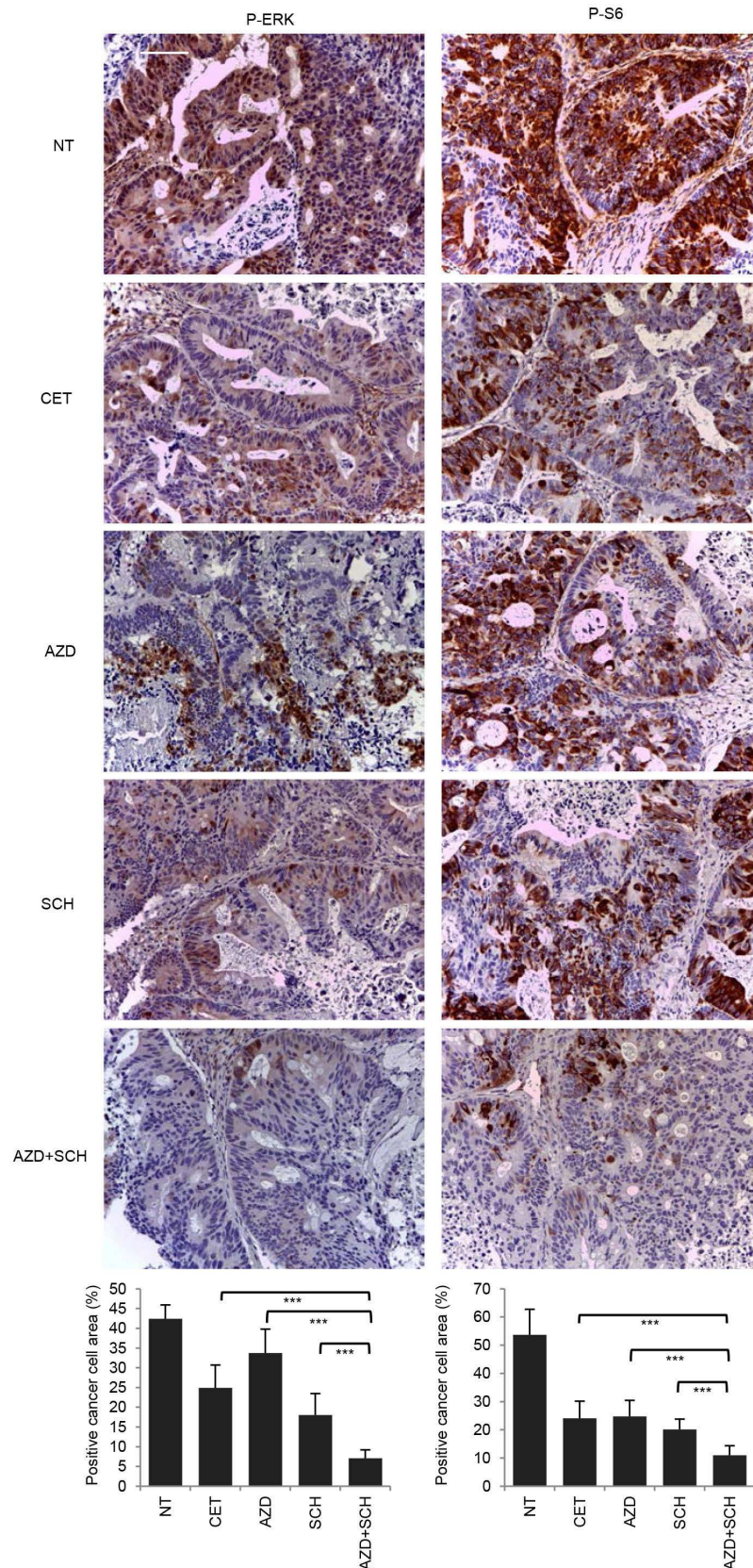
Extended Data Figure 6 | Signalling consequences of EGFR inhibition in EGFR mutant (V843I) CRC334. Immunohistochemistry with the indicated antibodies and morphometric quantitations of representative tumours at the

end of treatment. Results are the means \pm s.d. of five fields ($\times 40$) from two tumours for each experimental point ($n = 10$). Scale bar, 300 μ m. AFA, afatinib. ** $P < 0.01$; *** $P < 0.001$ by two-tailed Student's t -test.



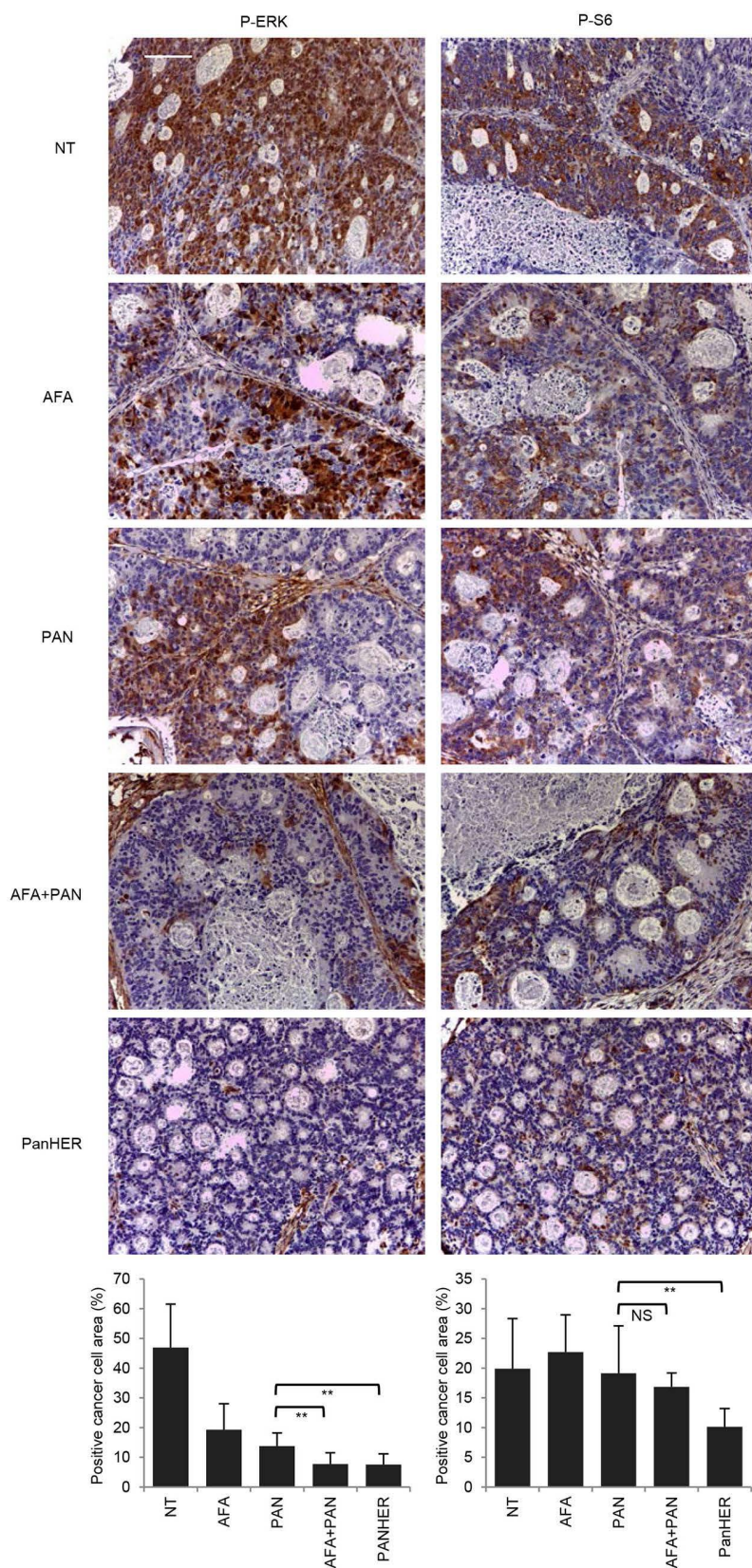
Extended Data Figure 7 | Signalling consequences of PDGFR inhibition in PDGFRA mutant (R981H) CRC525. Immunohistochemistry with the indicated antibodies and morphometric quantitations of representative tumours after acute treatment (4 h after imatinib and 24 h after cetuximab

administration). Results are the means \pm s.d. of five fields ($\times 40$) from two tumours for each experimental point ($n = 10$). Scale bar, 300 μ m. ** $P < 0.01$ by two-tailed Student's *t*-test.



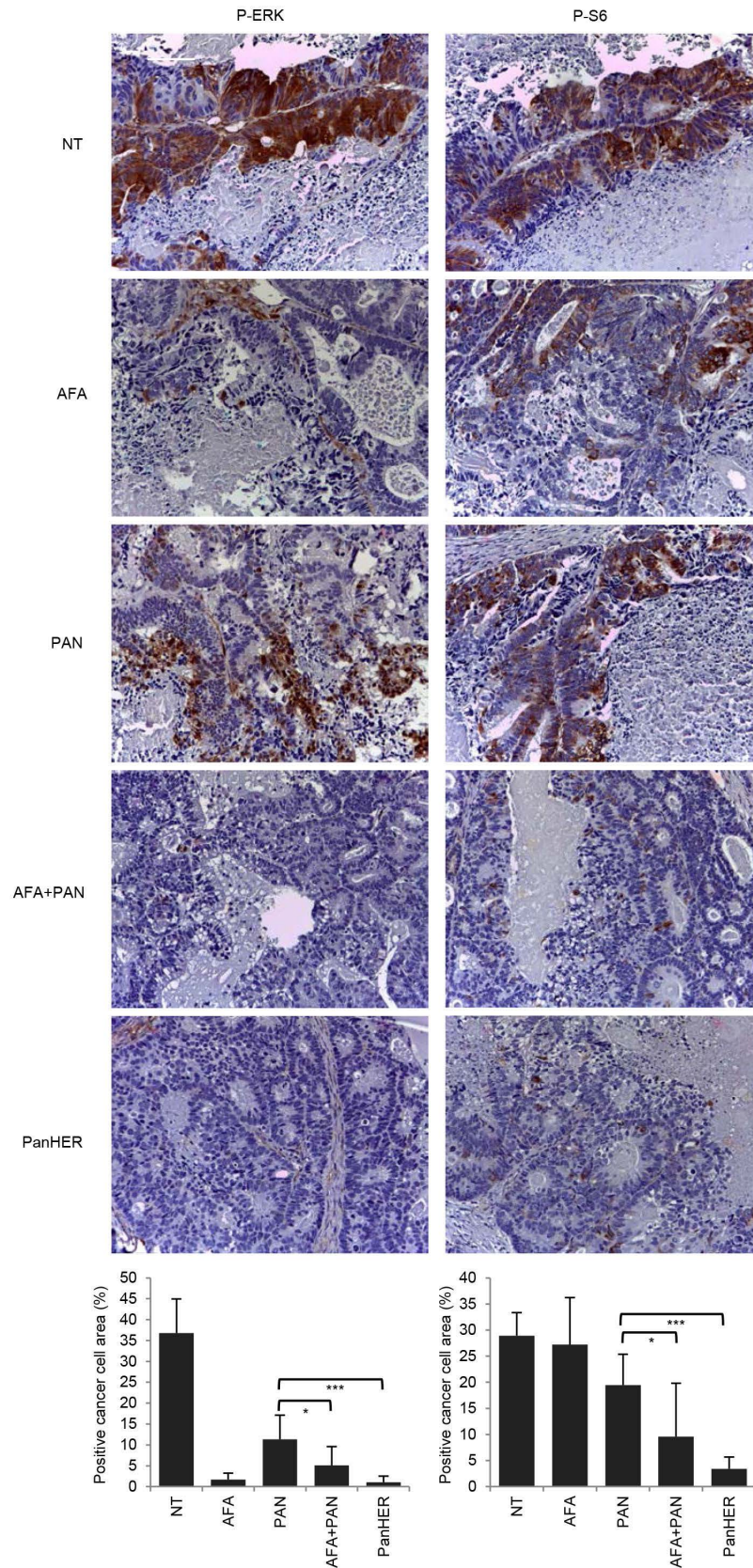
Extended Data Figure 8 | Signalling consequences of MEK1 inhibition in MAP2K1 mutant (K57KN) CRC343. Immunohistochemistry with the indicated antibodies and morphometric quantitations of representative tumours at the end of treatment. Results are the means \pm s.d. of five

fields ($\times 40$) from two tumours for each experimental point ($n = 10$). Scale bar, 300 μ m. AZD, AZD6244; SCH, SCH72984. *** $P < 0.001$ by two-tailed Student's t -test.



Extended Data Figure 9 | Signalling consequences of EGFR inhibition in EGFR mutant (G465E) CRC104. Immunohistochemistry with the indicated antibodies and morphometric quantitations of representative tumours at the end of treatment. Results are the means \pm s.d. of five fields

($\times 40$) from two tumours for each experimental point ($n = 10$). Scale bar, 300 μ m. PAN, panitumumab. NS, not significant; $**P < 0.01$ by two-tailed Student's t -test.



Extended Data Figure 10 | Signalling consequences of EGFR inhibition in EGFR mutant (G465R) CRC177. Immunohistochemistry with the indicated antibodies and morphometric quantitations of representative tumours at the

end of treatment. Results are the means \pm s.d. of five fields ($\times 40$) from two tumours for each experimental point ($n = 10$). Scale bar, 300 μ m. * $P < 0.05$; *** $P < 0.001$ by two-tailed Student's t -test.

Dilution of the cell cycle inhibitor Whi5 controls budding–yeast cell size

Kurt M. Schmoller¹, J. J. Turner¹, M. Kõivomägi¹ & Jan M. Skotheim¹

Cell size fundamentally affects all biosynthetic processes by determining the scale of organelles and influencing surface transport^{1,2}. Although extensive studies have identified many mutations affecting cell size, the molecular mechanisms underlying size control have remained elusive³. In the budding yeast *Saccharomyces cerevisiae*, size control occurs in G1 phase before Start, the point of irreversible commitment to cell division^{4,5}. It was previously thought that activity of the G1 cyclin Cln3 increased with cell size to trigger Start by initiating the inhibition of the transcriptional inhibitor Whi5 (refs 6–8). Here we show that although Cln3 concentration does modulate the rate at which cells pass Start, its synthesis increases in proportion to cell size so that its total concentration is nearly constant during pre-Start G1. Rather than increasing Cln3 activity, we identify decreasing Whi5 activity—due to the dilution of Whi5 by cell growth—as a molecular mechanism through which cell size controls proliferation. Whi5 is synthesized in S/G2/M phases of the cell cycle in a largely size-independent manner. This results in smaller daughter cells being born with higher Whi5 concentrations that extend their pre-Start G1 phase. Thus, at its most fundamental level, size control in budding yeast results from the differential scaling of Cln3 and Whi5 synthesis rates with cell size. More generally, our work shows that differential size-dependency of protein synthesis can provide an elegant mechanism to coordinate cellular functions with growth.

To control size, proliferating cells tie division to growth. However, the molecular mechanisms by which growth triggers division are poorly understood^{3,9,10}. In *S. cerevisiae*, which divides asymmetrically into a larger mother and smaller daughter cell, size control takes place in the first G1 phase of daughter cells^{5,11,12}. Progression through G1 is promoted by the upstream G1 cyclin Cln3 in complex with the cyclin-dependent kinase Cdk1. This Cln3–Cdk1 complex is thought to partly inactivate the transcriptional inhibitor Whi5 (refs 13, 14). Inactivation of Whi5 relieves inhibition of the transcription factor SBF, whose transcriptional activation completes a positive feedback loop committing the cell to division⁴. While its upstream position in the G1 regulatory network suggests that Cln3 is the trigger, its concentration does not clearly increase during G1 (refs 8, 15). Although Cln3 is a nuclear protein, the size of the yeast nucleus is proportional to cell size, so that the measured cellular concentrations reflect nuclear concentrations¹⁶. This leads to the question of why G1 progression is size-dependent when the putative trigger protein Cln3 does not increase in concentration.

Two prevailing models propose mechanisms to generate a size-dependent signal from the constant Cln3 concentration. One model proposes that the increasing number of Cln3 molecules is titrated against the fixed number of SBF-binding sites on the genome⁶. This DNA-titration serves to convert the constant Cln3 concentration to an increasing activity at its target sites on the genome. The other model proposes that Cln3 is retained at the endoplasmic reticulum and is rapidly released upon sufficient growth-dependent accumulation of the chaperone Ydj1 (refs 7, 17). Here we perform a series of experiments whose results are inconsistent with the two existing models.

Instead, we identify a new molecular mechanism for cell size control that does not require Cln3 activity increasing with cell size. Rather, we show that cell size promotes G1 progression by diluting the primary target of Cln3, Whi5 (Fig. 1a).

To determine how the G1 regulatory network implements size control, we first examined how the concentration of key regulators changes through G1. We grew cells using ethanol as the carbon source to generate small daughter cells subject to strong cell size control⁵. We restricted our attention to these daughter cells, and used time-lapse microscopy to measure the concentration of proteins tagged with the fluorescent protein mCitrine and expressed from the endogenous locus (Fig. 1b–g and Extended Data Fig. 1a). The concentration of wild-type (WT) Cln3 cannot be measured with this approach owing to its rapid and constitutive degradation. We therefore examined two mutants expressing stabilized proteins (*CLN3-11A* and *CLN3-1* (refs 18–20); Extended Data Fig. 1b). Consistent with previous bulk measurements of WT Cln3, Cln3-11A and Cln3-1 concentrations are constant through G1 (Fig. 1b). Moreover, we observed no changes in Cln3-11A localization (Extended Data Fig. 2). This is inconsistent with the Cln3 retention model, which predicts a rapid increase in nuclear Cln3 concentration in mid-G1 (ref. 7). Similarly, the concentrations of the key G1 regulators Swi4, Whi3, and Bck2 are nearly constant through G1 (Fig. 1c–e). In sharp contrast, we found that the concentration of the cell cycle inhibitor Whi5 strongly decreases through G1 (Fig. 1f and Extended Data Fig. 3a). This suggests that the dilution of the cell cycle inhibitor Whi5 is a size-dependent signal promoting cell cycle progression (Fig. 1a). Such an inhibitor-dilution model²¹ represents a qualitatively distinct mechanism of cell size control that does not require a size-dependent increase in Cln3-activity.

While inhibitor dilution explains how growth drives proliferation, it does not immediately explain why smaller-born cells grow more in G1. This would also require that smaller-born cells start G1 with a higher concentration of Whi5. Indeed, we found that the concentration of Whi5 at cell birth monotonically decreases with cell size, whereas the concentration of Cln3-11A and Cln3-1 at cell birth is independent of cell size (Fig. 2a, b). We confirmed that Whi5 is diluted in G1 using quantitative immunoblots (Fig. 2c). Finally, for a given birth size, diploid cells are born with a much higher Whi5 concentration (Fig. 2d and Extended Data Fig. 4a), which is consistent with the long-standing observation that cell size scales linearly with ploidy^{3,22}. Taken together, our data support a size control model in which all cells are born with a similar dose of Whi5, which they dilute by growth to progress through G1.

Inhibitor-dilution results in size-dependent cell cycle progression because smaller cells are born with higher Whi5 concentrations. To identify the origin of this size-dependent Whi5 concentration, we measured the rate of Whi5 synthesis throughout the cell cycle. We found that Whi5 is a stable protein, synthesized primarily during S/G2/M (Fig. 2e and Extended Data Figs 3b and 5a), consistent with previous mRNA measurements²³. Whi5 is differentially partitioned so that, following division, its concentration in daughter cells is consistently higher than in mother cells (Extended Data Fig. 6a, b). Critically,

¹Department of Biology, Stanford University, Stanford, California 94305, USA.

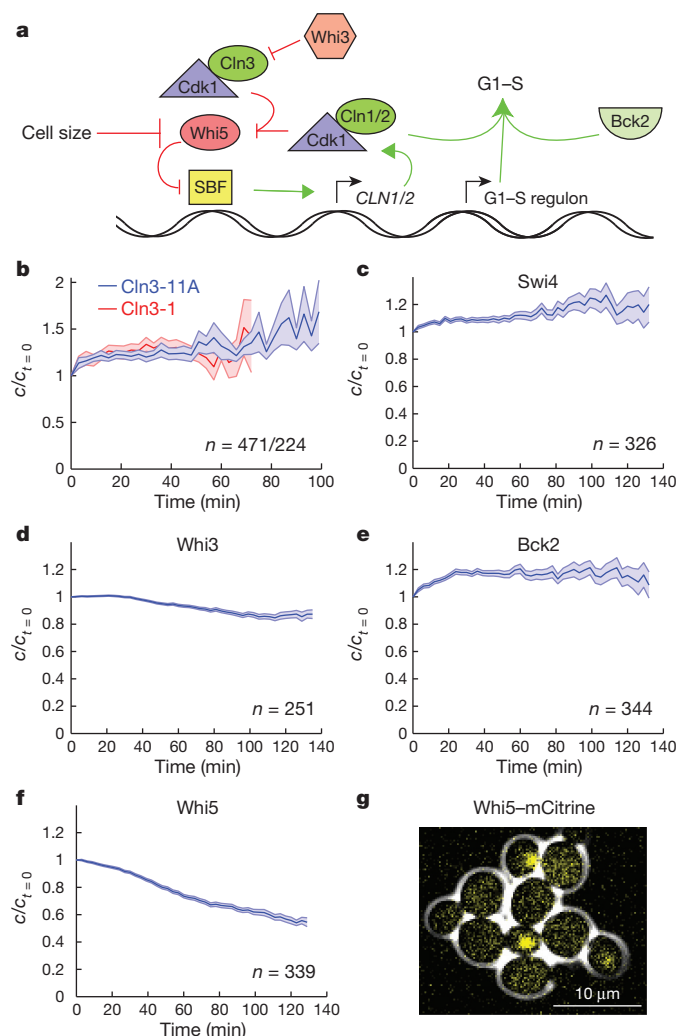


Figure 1 | The cell cycle inhibitor Whi5 is diluted by growth in G1.

a, Schematic of the G1/S regulatory network. **b–f**, Change in cellular protein concentration during G1 (mean \pm s.e.m.) for daughter cells expressing the indicated protein fused to mCitrine. Concentration was normalized to the concentration at cell birth ($t = 0$). **g**, Composite phase and fluorescence image of WHI5-mCitrine cells.

during S/G2/M, Whi5 is synthesized at a rate largely independent of cell size (Fig. 2e, f and Extended Data Figs 3b and 4b). Since S/G2/M duration weakly depends on mother cell size (Extended Data Fig. 6c), small and large cells produce similar amounts of Whi5. This results in larger budded cells having a lower Whi5 concentration just before division (Extended Data Fig. 6d). Since larger mother cells produce larger daughter cells (Extended Data Fig. 6e), this explains the inverse correlation between Whi5 concentration and cell size at birth, which is essential for the inhibitor-dilution size control model.

The size-independent synthesis rate of Whi5 during S/G2/M shows it is not limited by the general biosynthetic capacity of the cell, which increases with cell size. This contrasts with the expectation that transcriptional and translational outputs scale with cell size²⁴. Indeed, mCitrine-Cln3-11A synthesis in G1 is proportional to cell size (Fig. 2g and Extended Data Fig. 5b–c). Thus, the differential size scaling of Cln3 and Whi5 synthesis lies at the heart of cell size control in budding yeast.

Since ploidy is an important determinant of cell size, we decided to examine how it impacts the differential synthesis of Cln3 and Whi5. As expected for the majority of genes, whose synthesis is limited by the biosynthetic capacity of the cell, Cln3-11A synthesis in a diploid cell is comparable to the synthesis of a similarly sized haploid cell, despite

having two copies of mCitrine-CLN3-11A (Fig. 2g). Thus, in diploids the biosynthetic machinery is split between the two copies of the genome. Consistently, a hemizygous diploid synthesizes mCitrine-Cln3-11A protein at a much lower rate than a similarly sized haploid or homozygous diploid (Fig. 2g). In sharp contrast, Whi5-mCitrine synthesis is similar and size-independent in hemizygous diploid and haploid cells (Fig. 2f and Extended Data Fig. 4b). Moreover, a homozygous diploid produces Whi5 at approximately twice the rate, similar to a haploid with two copies of WHI5 (Fig. 2f and Extended Data Fig. 4b). Thus, the rate of Whi5 synthesis is determined by the number of copies of the gene and is independent of cell size and ploidy.

While the inhibitor-dilution model takes into account cell-to-cell variability in birth size, it does not yet include the fact that cells born the same size will vary in how much they grow before Start⁵ (Fig. 3a). For a population of similarly sized pre-Start cells, only a fraction will pass Start within the short time interval between movie frames. This allows us to define a rate as this fraction divided by the time interval (Fig. 3b; see Methods). In our inhibitor-dilution model, the rate at which cells pass Start is determined by the concentrations of Whi5 and Cln3. If Cln3 concentration is constant in pre-Start cells, the Whi5 concentration alone should predict the rate at which cells progress through Start. To test this model, we generated haploid strains containing one, two, and four copies of WHI5-mCitrine. We note these experiments are in a *bck2* Δ background, where Cln3 is essential²⁵. As expected, cells containing two and four copies of WHI5 produced proportionally more Whi5 protein, were larger, and exhibited a decreased size-dependent rate of progression through Start (Fig. 3b and Extended Data Fig. 4c, d). We note that these experiments were performed using cells expressing WT Cln3, which is suggested to be at constant concentration in G1 based on our measurements of Cln3-11A and Cln3-1. In complete agreement with an inhibitor-dilution model with a size-independent activator, the concentration of Whi5 alone predicts the rate at which cells progress through Start for all three strains (Fig. 3c). Consistently, the relationship between the rate of progression through Start and Whi5 concentration was not changed in *hcm1* Δ cells that lack a transcription factor promoting WHI5 expression²³ (Extended Data Fig. 7).

In our inhibitor-dilution model, Cln3 concentration determines the fraction of active Whi5. Thus, for a given Whi5 concentration, it should be possible to drive cell cycle entry by sufficiently increasing Cln3 concentration. To test this prediction, we constructed a *bck2* Δ strain with mCitrine-CLN3-11A under control of the methionine-regulated MET25 promoter. In this strain, repressing CLN3-11A expression arrests cells in G1, during which they continue to grow. Thus, by first arresting cells for varying durations and then inducing CLN3-11A for varying lengths of time, we were able to examine a wide range of cell sizes and Cln3 and Whi5 concentrations (Fig. 4a). We binned cells by size, which determines Whi5 concentration, and performed a logistic regression to determine the critical Cln3 concentration (pulse amplitude that results in half the cells budding; for example, Fig. 4b and Extended Data Fig. 8). Larger cells required lower Cln3-11A concentrations to enter the cell cycle (Extended Data Fig. 8d), consistent with previous results showing that larger G1 cells were more sensitive to Cln1 expression²⁶. Next, we used a strain that carries MET25pr-CLN3-11A and WHI5-mCitrine to measure the average Whi5 concentration as a function of cell size under the same arrest conditions (Extended Data Fig. 8e). The critical Cln3 concentration increases with Whi5 concentration as predicted by the Whi5-dilution model (Fig. 4c).

The Whi5-dilution model, unlike DNA-titration models, does not explicitly depend on the DNA content of the cell and predicts that the relationship between the critical Cln3 concentration and Whi5 concentration should be independent of ploidy. To test this, we repeated the same set of pulsing experiments using diploid strains and found a similar relationship between the critical Cln3 and Whi5 concentrations (Fig. 4c; $P > 0.05$). This is consistent with experiments showing

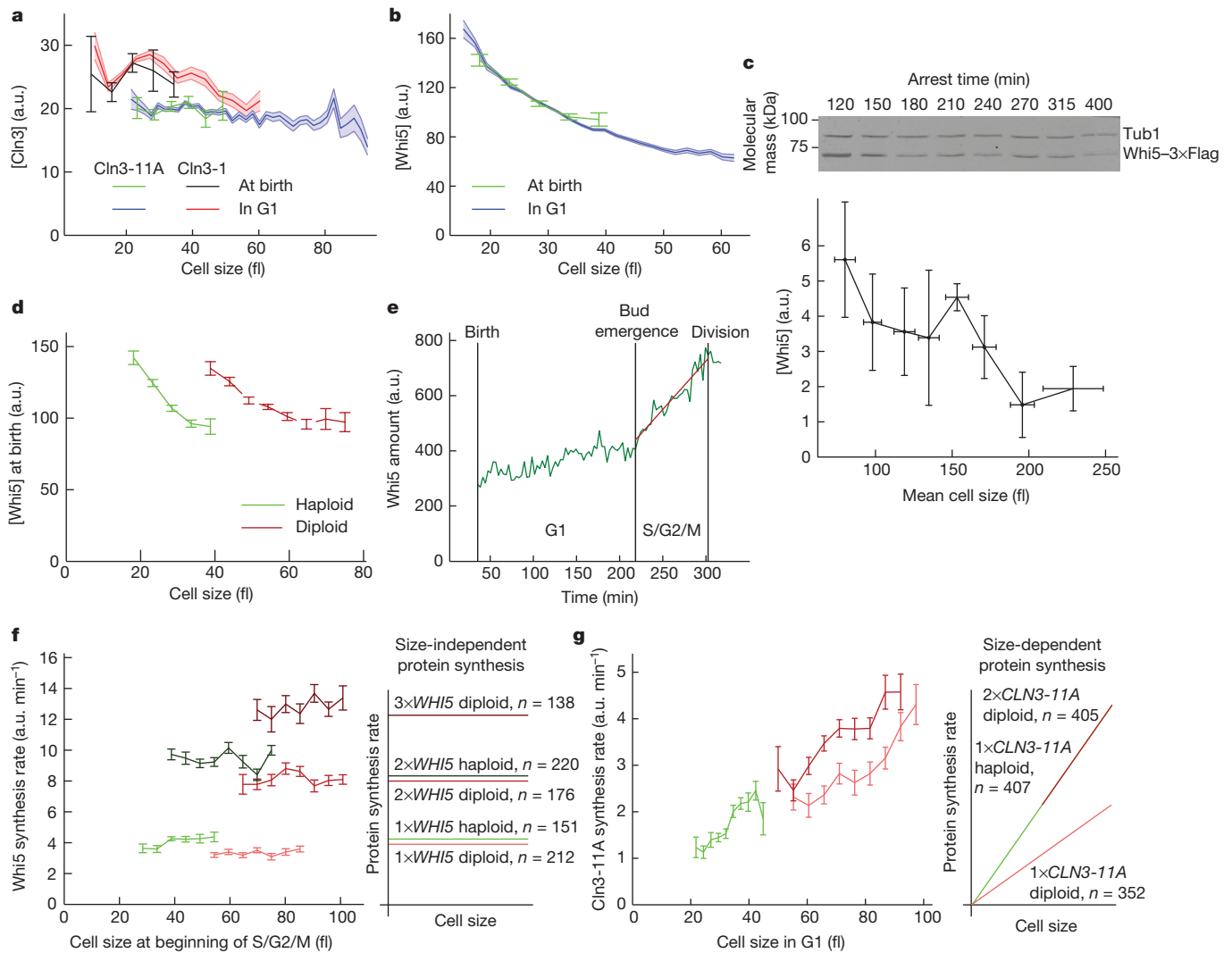


Figure 2 | Differential size-dependence of Cln3 and Whi5. **a, b**, Mean concentration of mCitrine-Cln3-11A ($n = 471$), mCitrine-Cln3-1 ($n = 234$) (**a**), and Whi5-mCitrine ($n = 339$) (**b**) as a function of cell size for daughter cells in G1. Shaded area, s.e.m.; bars, mean concentration and associated s.e.m. as a function of cell size at birth; a.u., arbitrary units. **c**, Top: representative immunoblot of cells arrested in G1 for increasing amounts of time; bottom: quantification of combined data from four independent time courses binned by mean population cell size (see Methods). Bars, means \pm s.d. Lanes were normalized by total protein content. Cells were grown on synthetic complete dextrose (SCD). **d**, Mean (s.e.m.) Whi5 concentration at cell birth is shown as a

function of cell size for haploid ($n = 339$) and diploid ($n = 385$) cells. **e**, Characteristic single-cell trace showing the total amount of Whi5-mCitrine in a haploid cell. The approximately linear increase in Whi5 during S/G2/M was fitted to determine the rate of synthesis. **f, g**, Mean rate of Whi5 (**f**) and Cln3-11A (**g**) synthesis as a function of cell size for each genotype indicated in the idealized schematics. Bars, s.e.m. The rate of Cln3 synthesis is proportional to cell size as indicated by its constant concentration in G1 shown in **a** (see Extended Data Fig. 5), whereas the Whi5 synthesis rate is largely size-independent. See Extended Data Fig. 4a, b for corresponding single-cell data.

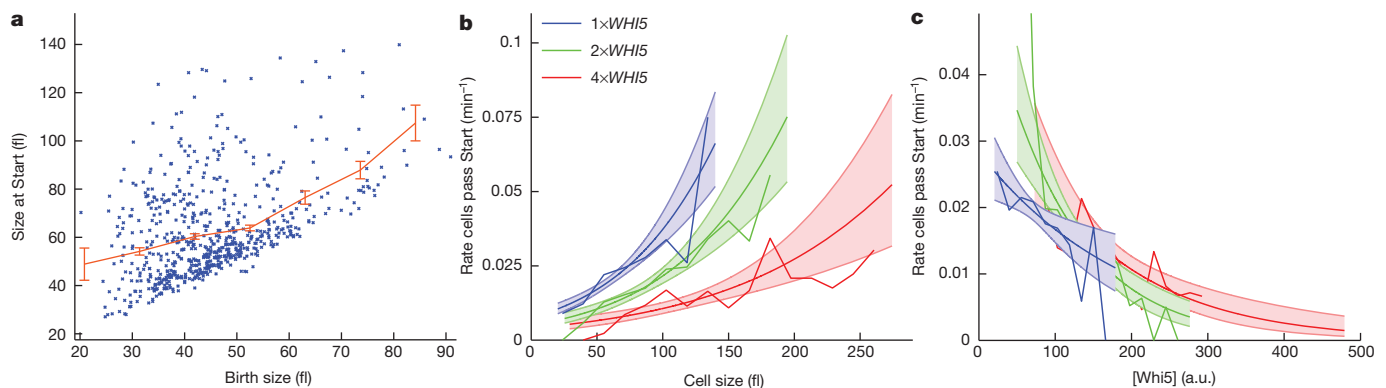


Figure 3 | Whi5 concentration determines the rate at which cells progress through Start. **a**, Size at Start, the point of commitment to cell division, as a function of birth size for haploid *bck2Δ* daughter cells ($n = 658$). Bars, mean and s.e.m. **b, c**, The rate at which daughter cells progress through Start is shown as a

function of cell size (**b**) and Whi5 concentration (**c**) for *bck2Δ* haploid cells with one (blue, $n = 658$), two (green, $n = 310$), or four (red, $n = 142$) copies of *WHI5*-mCitrine. Smooth lines are logistic regressions and the corresponding shaded areas denote 95% confidence intervals. Jagged lines connect means for binned data.

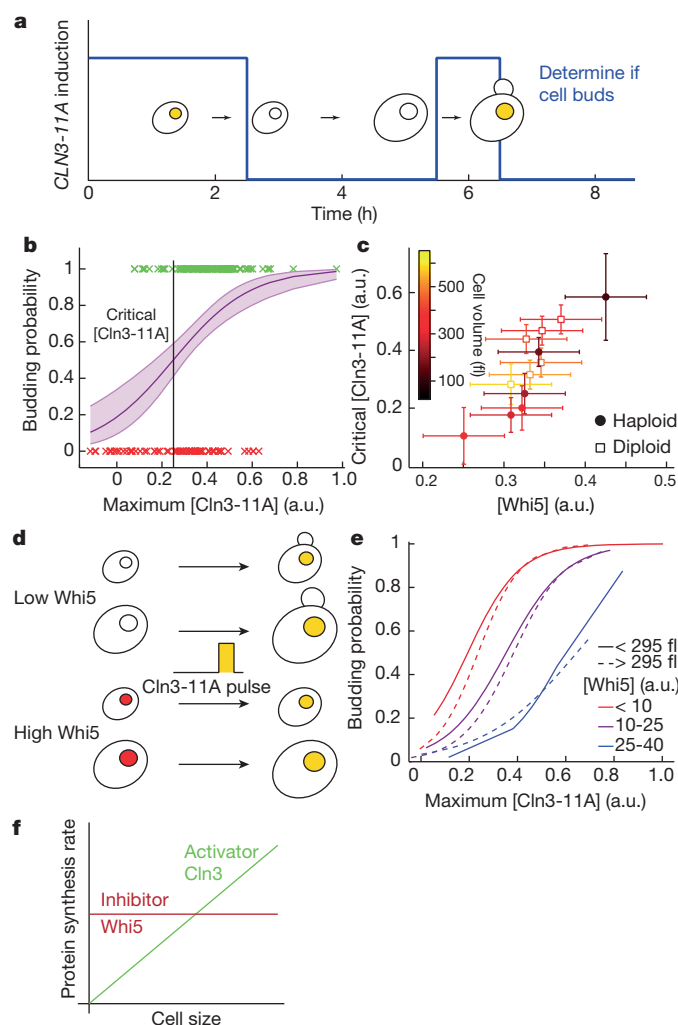


Figure 4 | Cln3 and Whi5 concentrations determine the rate at which cells pass Start irrespective of ploidy and cell size. **a**, Schematic of Cln3-pulse experiments. *MET25pr-mCitrine-CLN3-11A bck2Δ* cells are arrested in G1 for 2–4 h by addition of methionine to create G1 daughter cells of varying size. Following arrest, *mCitrine-CLN3-11A* expression is induced for varying amounts of time (0–60 min). **b**, Cells are binned by size and a logistic regression is then used to calculate the fraction of cells driven into the cell cycle as a function of the maximum Cln3-11A concentration produced by the exogenous pulse. Regression shown for 200–250 fl cells to determine the critical Cln3 concentration, where 50% of the cells bud. Single-cell data are marked green (budding) or red (not budding). **c**, The critical Cln3-11A concentration increases with Whi5 concentration, which was determined independently for each size bin (Extended Data Fig. 8). Haploid (filled circles, $n = 1195$) and diploid (open squares, $n = 405$) cells show a similar relationship between Whi5 and critical Cln3-11A concentrations ($P > 0.05$). **d**, To decouple cell size and Whi5 concentration, the pulse experiment is repeated with a strain expressing both *WHI5* and *CLN3-11A* from exogenously controlled promoters (*MET25pr-mCitrine-CLN3-11A LexApr-WHI5-mCherry bck2Δ*). Whi5 concentration and the duration of G1 arrest before the Cln3 pulse are varied. **e**, Data are displayed using two size and three Whi5 concentration bins. Higher Cln3 pulse amplitudes are needed to drive cells with higher Whi5 levels into the cell cycle, while no significant difference is observed for smaller and larger cells ($P > 0.5$, $n = 471$), which we note also have different DNA concentrations. **f**, Differential size-scaling of protein synthesis underlies budding yeast size control and provides a general mechanism to measure cell size independently of cell geometry.

that introducing heterologous DNA through yeast artificial chromosomes does not affect progression through G1 (ref. 27). However, we note that increased ploidy delays progression through S/G2/M, which

results in larger daughter cells and mean population size (Fig. 2f and Extended Data Fig. 9).

To determine that the relevant parameter for Start is Whi5 concentration, rather than cell size, we repeated the pulse experiments with a strain that carries a hormone-inducible promoter expressing a *WHI5-mCherry* allele in addition to the *MET25pr-mCitrine-CLN3-11A* allele (Fig. 4d). This allowed us to generate cells of different sizes containing similar Whi5 concentrations. As predicted, the probability of a cell passing Start increases with Cln3 concentration ($P < 10^{-5}$), decreases with Whi5 concentration ($P < 10^{-5}$), but is independent of cell size ($P > 0.5$) (Fig. 4e).

Taken together, our data support a new inhibitor-dilution model for size control in budding yeast. In this model, cell growth dilutes the cell cycle inhibitor Whi5 to drive progression through the cell cycle, whereas Cln3 concentration and activity remain constant. In this model, any regulation of Cln3 transcription²⁸, translation²⁹, or stability³⁰ that affects Cln3 concentration can be used to modulate cell size in different environmental conditions. While inhibitor dilution immediately leads to a growth requirement, this requirement is not necessarily different for larger- and smaller-born daughter cells, which is necessary for cell size control. For the inhibitor-dilution mechanism to control cell size it is crucial that the amount of Whi5 that cells are born with does not scale with size as we have shown here. In contrast, the cell cycle activator Cln3 is produced in proportion to size. This differential size-dependency of cell cycle activator and inhibitor synthesis constitutes the basis of size control in budding yeast (Fig. 4f).

Inhibitor-dilution is an elegant mechanism to control cell size independently of other aspects of cell geometry. In rod-shaped cells, such as fission yeast and bacteria, geometric mechanisms measuring lengths or surface areas can sensitively measure cell size because these metrics are directly proportional. However, such geometric measurements would perform poorly in near-spherical budding yeast, where intracellular lengths scale with the cube root of cell volume, so that a doubling of cell size results in only an ~25% increase in characteristic length. Geometric mechanisms are also unlikely to be applicable to more irregularly shaped metazoan cells. In contrast, inhibitor-dilution mechanisms can measure cell volume with no geometric constraints. All that is required is the differential size scaling of a cell cycle activator relative to an inhibitor. Because of the simplicity of this requirement, we anticipate the wide application of inhibitor-dilution mechanisms to control cell size.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 16 November 2014; accepted 14 July 2015.

Published online 21 September 2015.

- Goehring, N. W. & Hyman, A. A. Organelle growth control through limiting pools of cytoplasmic components. *Curr. Biol.* **22**, R330–R339 (2012).
- Chan, Y.-H. M. & Marshall, W. F. Scaling properties of cell and organelle size. *Organogenesis* **6**, 88–96 (2010).
- Turner, J. J., Ewald, J. C. & Skotheim, J. M. Cell size control in yeast. *Curr. Biol.* **22**, R350–R359 (2012).
- Donicic, A., Falleur-Fettig, M. & Skotheim, J. M. Distinct interactions select and maintain a specific cell fate. *Mol. Cell* **43**, 528–539 (2011).
- Di Talia, S., Skotheim, J. M., Bean, J. M., Siggia, E. D. & Cross, F. R. The effects of molecular noise and size control on variability in the budding yeast cell cycle. *Nature* **448**, 947–951 (2007).
- Wang, H., Carey, L. B., Cai, Y., Wijnen, H. & Futcher, B. Recruitment of Cln3 cyclin to promoters controls cell cycle entry via histone deacetylase and other targets. *PLoS Biol.* **7**, e1000189 (2009).
- Vergés, E., Colomina, N., Gari, E., Gallego, C. & Aldea, M. Cyclin Cln3 is retained at the ER and released by the J chaperone Ydj1 in late G1 to trigger cell cycle entry. *Mol. Cell* **26**, 649–662 (2007).
- Tyers, M., Tokiwa, G. & Futcher, B. Comparison of the *Saccharomyces cerevisiae* G1 cyclins: Cln3 may be an upstream activator of Cln1, Cln2 and other cyclins. *EMBO J.* **12**, 1955–1968 (1993).
- Lloyd, A. C. The regulation of cell size. *Cell* **154**, 1194–1205 (2013).
- Ginzberg, M. B., Kafri, R. & Kirschner, M. On being the right (cell) size. *Science* **348**, 1245075 (2015).

11. Di Talia, S. *et al.* Daughter-specific transcription factors regulate cell size control in budding yeast. *PLoS Biol.* **7**, e1000221 (2009).
12. Johnston, G. C., Pringle, J. R. & Hartwell, L. H. Coordination of growth with cell division in the yeast *Saccharomyces cerevisiae*. *Exp. Cell Res.* **105**, 79–98 (1977).
13. de Bruin, R. A. M., McDonald, W. H., Kalashnikova, T. I., Yates, J. & Wittenberg, C. Cln3 activates G1-specific transcription via phosphorylation of the SBF bound repressor Whi5. *Cell* **117**, 887–898 (2004).
14. Costanzo, M. *et al.* CDK activity antagonizes Whi5, an inhibitor of G1/S transcription in yeast. *Cell* **117**, 899–913 (2004).
15. Landry, B. D., Doyle, J. P., Toczyski, D. P. & Benanti, J. A. F-box protein specificity for G1 cyclins is dictated by subcellular localization. *PLoS Genet.* **8**, e1002851 (2012).
16. Jorgensen, P. *et al.* The size of the nucleus increases as yeast cells grow. *Mol. Biol. Cell* **18**, 3523–3532 (2014).
17. Yahya, G., Parisi, E., Flores, A., Gallego, C. & Aldea, M. A Whi7-anchored loop controls the G1 Cdk-cyclin complex at Start. *Mol. Cell* **53**, 115–126 (2014).
18. Bhaduri, S. & Pryciak, P. M. Cyclin-specific docking motifs promote phosphorylation of yeast signaling proteins by G1/S Cdk complexes. *Curr. Biol.* **21**, 1615–1623 (2011).
19. Liu, X. *et al.* Reliable cell cycle commitment in budding yeast is ensured by signal integration. *eLife* **4**, e03977 (2015).
20. Tyers, M., Tokiwa, G., Nash, R. & Futcher, B. The Cln3-Cdc28 kinase complex of *S. cerevisiae* is regulated by proteolysis and phosphorylation. *EMBO J.* **11**, 1773–1784 (1992).
21. Fantes, P. A., Grant, W. D., Pritchard, R. H., Sudbery, P. E. & Wheals, A. E. The regulation of cell size and the control of mitosis. *J. Theor. Biol.* **50**, 213–244 (1975).
22. Wu, C.-Y., Rolfe, P. A., Gifford, D. K. & Fink, G. R. Control of Transcription by Cell Size. *PLoS Biol.* **8**, e1000523 (2010).
23. Pramila, T., Wu, W., Miles, S., Noble, W. S. & Breeden, L. L. The Forkhead transcription factor Hcm1 regulates chromosome segregation genes and fills the S-phase gap in the transcriptional circuitry of the cell cycle. *Genes Dev.* **20**, 2266–2278 (2006).
24. Marguerat, S. & Bähler, J. Coordinating genome expression with cell size. *Trends Genet.* **28**, 560–565 (2012).
25. Epstein, C. B. & Cross, F. R. Genes that can bypass the CLN requirement for *Saccharomyces cerevisiae* cell cycle START. *Mol. Cell. Biol.* **14**, 2041–2047 (1994).
26. Schneider, B. L. *et al.* Growth rate and cell size modulate the synthesis of, and requirement for, G1-phase cyclins at Start. *Mol. Cell. Biol.* **24**, 10802–10813 (2004).
27. Thorburn, R. R. *et al.* Aneuploid yeast strains exhibit defects in cell growth and passage through START. *Mol. Biol. Cell* **24**, 1274–1289 (2013).
28. Shi, L. & Tu, B. P. Acetyl-CoA induces transcription of the key G1 cyclin CLN3 to promote entry into the cell division cycle in *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA* **110**, 7318–7323 (2013).
29. Polymenis, M. & Schmidt, E. V. Coupling of cell division to cell growth by translational control of the G1 cyclin CLN3 in yeast. *Genes Dev.* **11**, 2522–2531 (1997).
30. Menoyo, S. *et al.* Phosphate-activated cyclin-dependent kinase stabilizes G1 cyclin to trigger cell cycle entry. *Mol. Cell. Biol.* **33**, 1273–1284 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank O. Atay and J. Feldman for reagents, R. de Bruin, A. Gladfelter, M. Cyert, and M. Loog for comments on the manuscript, the Burroughs Wellcome Fund (CASI), the National Science Foundation (CAREER), National Institutes of Health training grant GM007276 (to J.J.T.), and Human Frontier Science Program (postdoctoral fellowships to K.M.S. and M.K.) for funding.

Author Contributions All authors designed experiments and wrote the manuscript; K.M.S., J.T. and M.K. performed experiments.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.M.S. (skoheim@stanford.edu).

METHODS

Imaging and image analysis. All experiments were performed using a CellASIC microfluidics device with Y04C plates. A Zeiss Observer Z1 microscope with an automated stage using a plan-apo 63X/1.4NA oil immersion objective was used to take images every 3 min. Focusing was performed using automated Definite Focus hardware. Strains expressing mCitrine fusion proteins were exposed for 400 ms using the Colibri 505 LED module at 25% power. Whi5-mCherry was imaged by exposure for 500 ms using the Colibri 540-80 LED module at 50% power. Under these illumination conditions, we did not observe detectable photobleaching (Extended Data Fig. 10a). Cell and nuclear segmentation and quantification of fluorescence signals was performed as in ref. 31. We subtracted the size-dependent autofluorescence signal as determined from comparable experiments with unlabelled strains (Extended Data Fig. 10b, c) to measure the total fluorescence intensity in single cells from fluorescent-fusion proteins. Total fluorescence intensity is proportional to the protein amount. To determine protein concentration, we calculate cell volume from the phase image segmentation by assuming rotational symmetry around the major axis in the x - y plane. We confirmed that localization of Whi5 did not significantly affect concentration measurements (Extended Data Fig. 3c, d).

Experimental design and statistical analysis. All data shown were obtained from at least two independent experiments. For each experiment, we collected data from 10 to 20 imaging positions. Comparison of biological replicates allowed us to assay for systematic errors. This resulted in sufficient data for our statistical comparisons. No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Cell-cycle phase analysis. After automated segmentation, we manually annotated pedigrees and determined time points of cell birth, bud emergence, and cytokinesis from phase images. We only included daughter cells that were born during the experiment in our analyses. For cells shown in Figs 1 and 2, we estimated G1 using cytokinesis and bud emergence. Whi5 enters the nucleus ~9 min before cytokinesis⁵, which can also be detected by visual inspection of phase images. For Fig. 3, we estimated the time of Start, the point of commitment to cell division beyond which haploid cells no longer arrest in response to an abrupt exposure to a high concentration of mating pheromone. Start in haploid cells corresponds to the point where ~50% of the peak nuclear Whi5 has been removed from the nucleus⁴. Start takes place ~12 min before the full exit of Whi5-mCitrine from the nucleus in cells growing on 2% glycerol and 1% ethanol. For $2 \times$ WHI5-mCitrine and $4 \times$ WHI5-mCitrine strains, we used 12 min before full Whi5 exit as a proxy for Start.

Whi5 and Cln3 synthesis rate estimates. To estimate Whi5 synthesis rates, we analysed time series for the amount of Whi5-mCitrine in the S/G2/M phase of the cell cycle (the budded phase). This phase was determined as described above. We fitted a line to the data and took the slope as an estimate of the Whi5 synthesis rate because Whi5 is a highly stable protein (Extended Data Fig. 5a). Synthesis rate estimates for individual cells are plotted against the cell size at the time of bud emergence (Extended Data Fig. 4b); mean values for size-binned data are shown in Fig. 2f. We used a similar method to estimate the much lower Whi5 synthesis rate during G1 (Extended Data Fig. 3b).

To estimate Cln3 synthesis rates during G1, we analysed time series for individual cells expressing mCitrine-Cln3-11A. For each cell, we estimated the Cln3-11A synthesis rate, k , over a 30-min interval in which we assumed it was constant. This is valid because a cell growing on glycerol/ethanol changes little in size during a 30-min interval. We excluded cells with G1 durations shorter than 30 min. We assumed protein degradation can be characterized by an exponential decay constant $\tau \sim 83$ min that we independently measured (Extended Data Fig. 5c). We take N to be the number of Cln3-11A molecules, so that $dN/dt = k - \frac{1}{\tau}N$, which can be solved to yield the number of Cln3-11A molecules

$$N = N_0 e^{-t/\tau} + k\tau \left(1 - e^{-t/\tau}\right).$$

Here, N_0 denotes the amount of Cln3 at the

beginning of the interval. N_0 and k are then determined from fitting this equation to the data. The synthesis estimate k is then plotted against the cell size at the beginning of the interval (Fig. 2g).

Estimate of rate at which cells pass Start. The Start transition is a highly stochastic process⁵, which means that cells born at the same size will vary in how much time they spend growing in pre-Start G1. Thus, for a population of similarly sized pre-Start cells, only a fraction will pass Start within a given time interval. To quantify this phenomenon, we calculate the fraction of pre-Start cells within a size interval that pass Start within one frame of our movie (3 min). Thus, we define the rate at which cells pass Start as a function of cell size as the fraction of cells within the size bin that passed Start divided by the time interval between movie frames. Similarly, we can also define the rate at which cells pass Start as a function of Whi5 concentration by grouping cells by Whi5 concentration rather than cell

size. We note that the Start transition is defined in the cell-cycle phase analysis section above based on ref. 4. This analysis, based on binning cells by size or Whi5 concentration, was used to obtain the jagged lines in Fig. 3b, c. The smooth curves, and associated 95% confidence intervals, were obtained by logistic regression of the unbinned data set as follows. For each frame, a pre-Start cell is described by three numbers: cell size, Whi5 concentration, and whether or not that cell passed Start in the next 3 min (=1 if the cell passed Start; =0 otherwise). Data from all time points for all pre-Start daughter cells were pooled into a large matrix with three columns. We then performed a logistic regression using the MATLAB function `glmfit` to estimate the probability of a cell passing Start as a function of either cell size or Whi5 concentration.

We note that Fig. 3c shows that the instantaneous probability for a cell to pass Start is determined by the Whi5 concentration, not volume. However, this does not necessarily mean that cells with different Whi5 copy number need to reach the same Whi5 concentration. Consider two cells born with different Whi5 concentrations. The one with the higher Whi5 concentration has to grow for a certain amount of time to reach the initial Whi5 concentration of the second cell. Since the probability for passing Start is always non-zero, there is a certain chance that the cell enters the cell cycle even before reaching the birth concentration of the one with lower Whi5 concentration. Thus, on average, cells born with higher initial Whi5 concentrations will pass Start at higher Whi5 concentrations. Adding an extra copy of Whi5 does result in an increase of cell size, however, cell size is not doubled. Thus, $2 \times$ WHI5 cells are on average born with higher Whi5 concentrations, but also pass Start at higher Whi5 concentrations, even though the instantaneous probability of passing Start as a function of Whi5 is the same as in WT cells.

We also note that the series of experiments described in Figs 3 and 4 were performed in a *bck2Δ* background. In the absence of Cln3, Bck2 drives large cells into the cell cycle²⁵. However, since Bck2 concentration is constant through G1 (Fig. 1e), and the targets of this transcriptional regulator extend across the entire cell cycle^{32,33}, we decided to focus exclusively on the Cln3-Whi5 mechanism, which is specific for G1 progression, and performed subsequent analyses in a *bck2Δ* background.

Cln3 pulse experiments. For the experiments shown in Fig. 4, cells were grown on media lacking methionine, SCD-Met (*MET25pr-mCitrine-CLN3-11A* on). After 150 min of growth in the microfluidic device, $10 \times$ methionine was added to arrest cells in G1 (*MET25pr-mCitrine-CLN3-11A* off). After varying arrest times (2, 3, or 4 h for haploids, 3 h for diploids) methionine was removed for 0, 30, 40, 50, or 60 min (20, 30, 40, or 50 min for diploids) to induce a pulse of mCitrine-Cln3-11A expression. For daughter cells born during the experiment, we determined whether the cell budded during a 2 h time window following the onset of *CLN3-11A* induction. In addition, we measured the maximum Cln3-11A concentration. Cell size was measured at the time of maximum Cln3-11A concentration. For each strain, we then pooled all the data (22 independent experiments for haploids, 6 for diploids) and binned cells according to their size. For each size bin, we used a logistic regression to calculate the critical mCitrine-Cln3-11A concentration where 50% of the cells bud. To determine the median Whi5 concentration as a function of cell size we arrested *MET25pr-CLN3-11A bck2Δ WHI5-mCitrine* cells and measured Whi5 concentration as a function of cell size during the arrest (Extended Data Fig. 8). Error bars for Fig. 4c were calculated as the maximum of the 95% confidence interval of the logistic regression and the estimated experimental error due to variation in fluorescence intensity from experiment to experiment. We used a linear regression model to test whether ploidy affects the relationship between Whi5 and critical Cln3-11A concentrations. For the experiments shown in Fig. 4e and Extended Data Fig. 8, we used a hormone-inducible *LexApr*³⁴ to express a *Whi5-mCherry* allele to decouple cell size and Whi5 concentration. We induced Whi5 by addition of β -oestradiol (30–100 nM) for at least 6 h, and removed β -oestradiol before the experiment. This allowed us to generate cells of varying size and Whi5 concentration by varying the induction level of Whi5 and the duration of G1 arrest (2, 2.5, or 3 h; 18 independent experiments) before the Cln3-11A pulse induction in SCD-Met.

Growth conditions. For microscopy-based experiments, yeast were grown in synthetic complete media with 2% glycerol and 1% ethanol except for the pulse experiments shown in Fig. 4, where yeast were grown on synthetic complete media with 2% glucose. Before an imaging experiment, cells were grown to an absorbance <0.1 after which they were sonicated for ~5 s at 3 W intensity. For quantitative immunoblots, cells were grown on synthetic complete 2% galactose 2% raffinose overnight before being arrested in synthetic complete 2% glucose.

Strains and plasmids. All strains were congenic with W303 (see Supplementary Table 1), and were constructed using standard methods. See Supplementary Table 2 for plasmid list. To enable fluorescent Cln3 detection in live cells, it was necessary to use a stabilized variant of the protein. Stabilizing Cln3 by removing its degradation-inducing phosphosites¹⁸ (Cln3-10A) allowed direct observation of Cln3-10A-mCitrine. However, induction of this stabilized variant with the *MET25*

promoter resulted in severe cytokinesis defects. We therefore added an additional mutation (R108A), which was reported to increase protein stability, but reduce the ability of Cln3 to drive cell cycle progression¹⁹. Also, R108A has been examined in the context of the double mutant K106A R108A (see variant Cln3-A10 in ref. 35) that increased steady state protein levels, but was less able to rescue a *cln1Δ cln2Δ cln3Δ bck2Δ* strain. Cln3-A10 was also reported to decrease the interaction with Cdc28 in ref. 36 (where it is known as the *Cln3-A13* mutation). Combining the ten stabilizing alanine mutations from ref. 18 with the R108A mutation from ref. 19 resulted in a stabilized, less active Cln3 protein (Extended Data Fig. 1b), which we refer to as Cln3-11A, whose concentration and amount were measurable in single cells without disrupting cytokinesis. For Figs 1b and 2a, we measured the concentration of Cln3 protein expressed from a previously characterized stabilized C-terminal truncation allele *CLN3-1* (refs 20, 37).

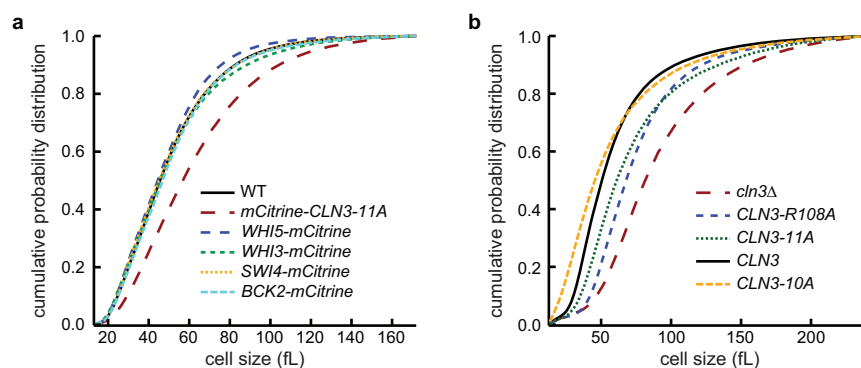
Quantitative Whi5 immunoblot. Strain JTY6 (*cln1Δcln2Δcln3Δ GAL1pr-CLN1 WHI5-3 × Flag*) was generated using plasmid pMK15 digested with MluI (New England BioLabs). Cells were grown at 30°C to mid-log phase (absorbance at 600 nm = 0.2) in synthetic complete media with 2% galactose and 2% raffinose before being washed once and resuspended in an equal volume of synthetic complete with 2% glucose and incubated for 120 min at 30°C. Ten-millilitre samples were removed every 30 min for 150 min, and then every 45 min for 90 min thereafter. Samples were pelleted and frozen in liquid nitrogen. Concurrently, 1 ml samples were removed, sonicated, and analysed with a Coulter Z2 cell counter to measure cell size distributions with size cutoffs set at 30 and 500 fl.

Frozen cell pellets were thawed on ice, resuspended in 200 µl urea lysis buffer (20 mM Tris•Cl pH 7.5, 7 M urea, 2 M thiourea, 65 mM CHAPS, 65 mM DTT, 50 mM NaF, 100 mM β-glycerophosphate, 1 mM NaVO₃, 1 mM PMSF), and homogenized for 40 s at 4°C in a FastPrep homogenizer (MP Biomedicals) using an equal volume of 0.5 mm diameter ceramic beads. Cell lysates were transferred to fresh microfuge tubes by puncturing the bottom of the tubes used for lysis, placing them in fresh tubes, and centrifuging for a few seconds at low speed. Lysates were then cleared by centrifuging at 17,000g for 10 min. Total protein concentration in the lysates was determined by Bradford analysis, and samples were diluted to a maximum volume of 12 µl in urea buffer, of which 10 µl were mixed with 5 µl 6 × Laemmli sample buffer and run on a 12% (29:1) polyacrylamide gel. Gels were cut to include only the relevant molecular weight range, and

proteins from all gels were transferred to a nitrocellulose membrane using program 8 for 7 min on an iBlot dry transfer device (Thermo Fisher Scientific). Membranes were blocked in Licor Odyssey blocking buffer (TBS; 927-50010) for 30 min at room temperature (~23°C). Membranes were incubated with 1:1,000 M2 mouse monoclonal anti-Flag (Sigma F1804) and 1:5,000 rat monoclonal anti-tubulin YOL1/34 (Abcam ab6161) diluted in Licor Odyssey blocking buffer + 0.2% Tween-20 for 60 min, and they were washed 1 × 15 min and 3 × 5 min in TBS + 0.1% Tween-20. Membranes were then incubated in 1:15,000 goat anti-mouse conjugated to Alexa Fluor 680 (Thermo Fisher Scientific A-21058) and goat anti-rat conjugated to Licor IRDye 800CW (Licor, 925-32219) in the same buffer as for primary antibodies and washed as before. Membranes were imaged in a Licor Odyssey CLX-0670.

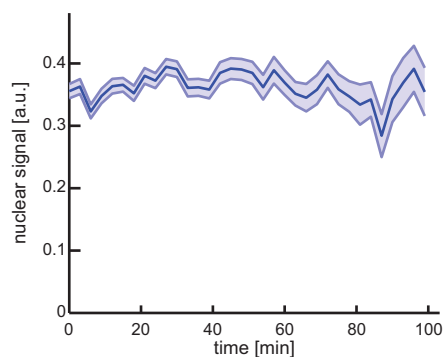
Immunoblot images were analysed by manually specifying band boundaries and measuring total intensity using Licor Image Studio Light. Background regions for each lane were also specified manually. These values were used to generate background-subtracted values, which were analysed with respect to mean population size using R.

31. Doncic, A., Eser, U., Atay, O. & Skotheim, J. M. An algorithm to automate yeast segmentation and tracking. *PLoS One* **8**, e57970 (2013).
32. Bastajian, N., Friesen, H. & Andrews, B. J. Bck2 acts through the MADS box protein Mcm1 to activate cell-cycle-regulated genes in budding yeast. *PLoS Genet.* **9**, e1003507 (2013).
33. Ferrezuelo, F., Aldea, M. & Futcher, B. Bck2 is a phase-independent activator of cell cycle-regulated genes in yeast. *Cell Cycle* **8**, 239–252 (2009).
34. Ottoz, D. S. M., Rudolf, F. & Stelling, J. Inducible, tightly regulated and growth condition-independent transcription factor in *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **42**, e130 (2015).
35. Miller, M. E., Cross, F. R., Groeger, A. L. & Jameson, K. L. Identification of novel and conserved functional and structural elements of the G1 cyclin Cln3 important for interactions with the CDK Cdc28 in *Saccharomyces cerevisiae*. *Yeast* **22**, 1021–1036 (2005).
36. Miller, M. E. & Cross, F. R. Mechanisms controlling subcellular localization of the G(1) cyclins Cln2p and Cln3p in budding yeast. *Mol. Cell. Biol.* **21**, 6292–6311 (2001).
37. Nash, R., Tokiwa, G., Anand, S., Erickson, K. & Futcher, A. B. The WHI1+ gene of *Saccharomyces cerevisiae* tethers cell division to cell size and is a cyclin homolog. *EMBO J.* **7**, 4335–4346 (1988).

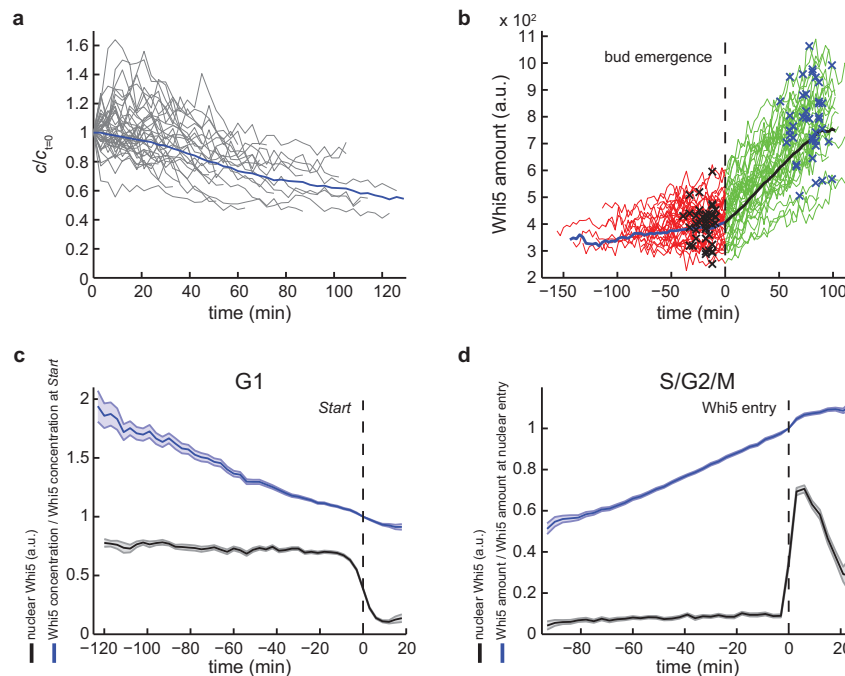


Extended Data Figure 1 | Size distributions of strains expressing mCitrine fusion proteins or *CLN3* mutant alleles. **a**, Cell size distributions were measured in a Coulter counter for five strains expressing the indicated mCitrine fusion proteins from the endogenous locus and a WT control. These five strains were used in Fig. 1. All strains were grown on synthetic complete 2%

glycerol, 1% ethanol. **b**, Size distributions measured using a Coulter counter for *cln3Δ* cells expressing *CLN3* alleles from a *CLN3* promoter integrated at the *URA3* locus. See Methods for description of *CLN3* mutant alleles. Cells were grown on synthetic complete 2% glucose.



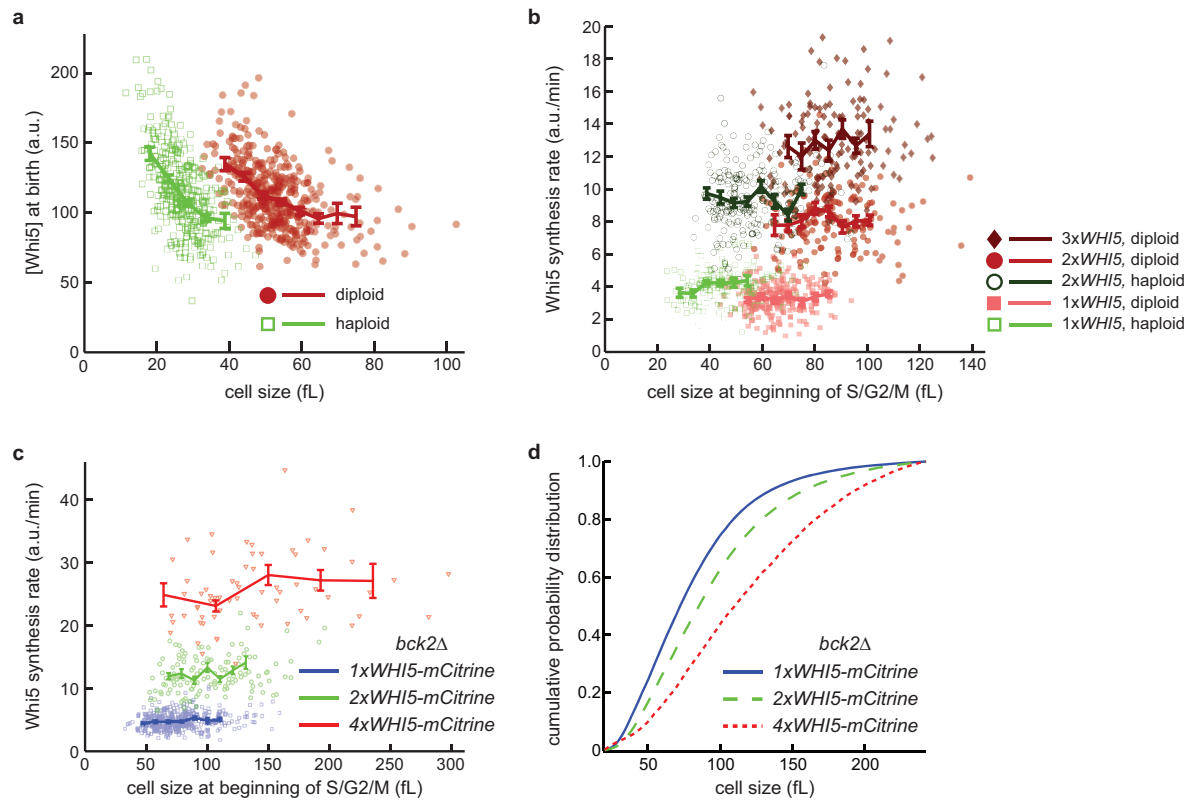
Extended Data Figure 2 | mCitrine-Cln3-11A is consistently nuclear during G1. We see no evidence of a rapid re-localization of Cln3-11A into the nucleus at mid-G1 ($n = 471$). Nuclear signal measured and nucleus segmented as described in ref. 31. Thick line denotes mean; shaded area denotes s.e.m.



Extended Data Figure 3 | Single-cell analysis of Whi5 dilution and synthesis.

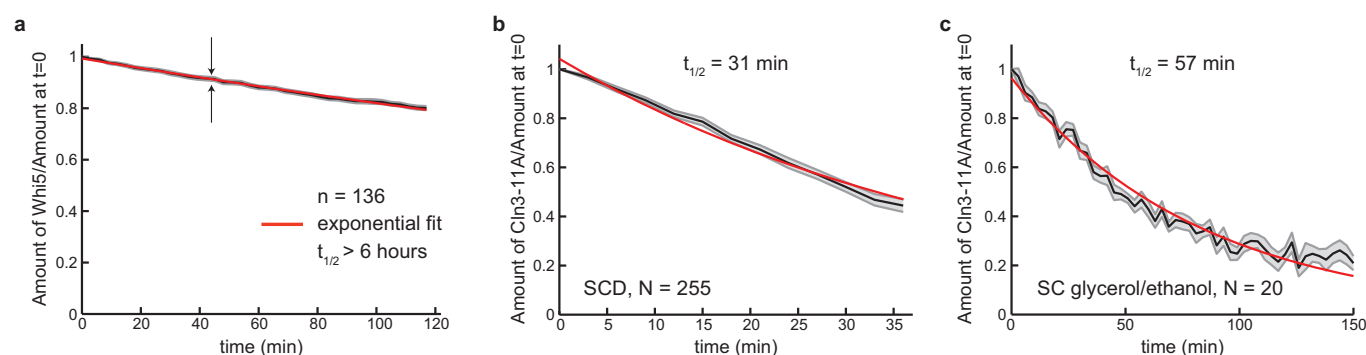
a, We randomly selected 40 out of 339 single-cell traces that correspond to the data shown in Fig. 1f for display here. The relative change of Whi5 concentration during G1 is shown in grey (thin lines). Blue thick line shows the mean of all 339 cells. **b**, We randomly selected 40 out of 147 single-cell traces of Whi5 amount for display. Traces are aligned by bud emergence ($t = 0$). G1 (cell birth to bud emergence) is shown in red (mean of all 147 cells is shown in blue). S/G2/M (bud emergence to cytokinesis) is shown in green (mean is shown in black). Black crosses denote time points of full nuclear Whi5 exit, blue crosses denote time points of full nuclear Whi5 re-entry. The rate of synthesis of Whi5 in S/G2/M phase is 6.6-fold higher than in G1 phase. Eighty-nine per cent of total Whi5 in this experiment is synthesized in S/G2/M. **c, d**, Control for the effect of Whi5 localization on concentration measurements. Rapid relocalization of Whi5 at Start (**c**), and just before

cytokinesis (**d**), does not affect concentration measurements. **c**, Mean relative change of cellular Whi5 concentration during G1 aligned by Start (50% nuclear Whi5 exit as determined from a logistic fit to the nuclear signal) and s.e.m. are shown in blue. The corresponding nuclear Whi5 signal is shown in black (mean and s.e.m.; nuclear signal measured and nucleus segmented as described in ref. 31); $n = 320$. **d**, Mean relative change of cellular Whi5 amount during S/G2/M in mother-bud pairs aligned by the time point of 50% Whi5 entry into the nucleus (as determined from a logistic fit to the nuclear signal) and s.e.m. are shown in blue. The corresponding nuclear Whi5 signal is shown in black (mean and s.e.m.; nuclear signal measured and nucleus segmented as described in ref. 31); $n = 133$. Cells express Whi5-mCitrine from the endogenous locus. Cells were grown on synthetic complete 2% glycerol, 1% ethanol.



Extended Data Figure 4 | Whi5 concentration and synthesis rate. Single-cell data corresponding to Fig. 2d, f. **a**, Whi5 concentration at cell birth is shown as a function of cell size for individual haploid ($n = 339$) and diploid ($n = 385$) cells. **b**, The rate of Whi5 synthesis as a function of cell size for each genotype as indicated. **c**, The rate of Whi5 synthesis as a function of cell size for

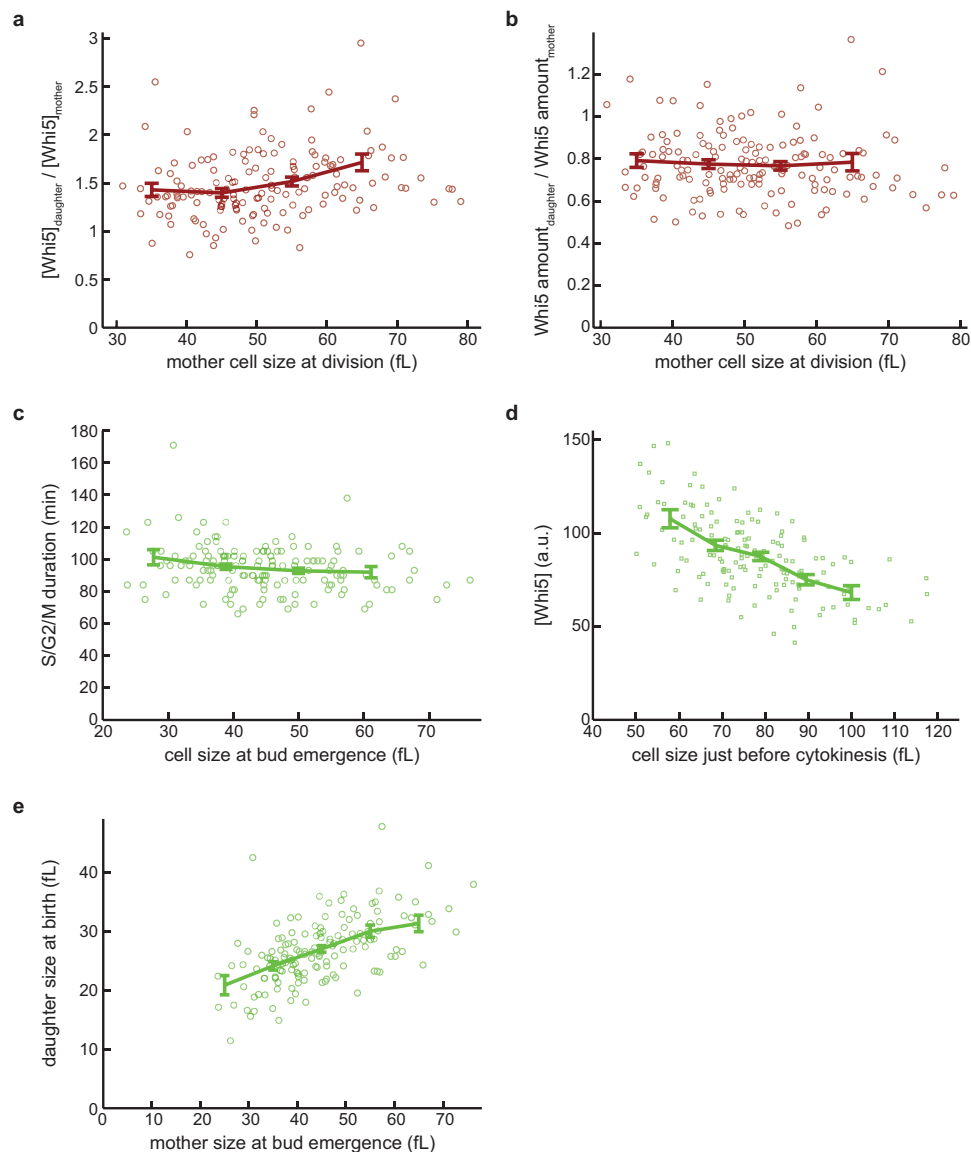
bck2Δ strains expressing one, two, or four copies of *WHI5-mCitrine* ($n = 353$, 129 and 66, respectively). Bars denote means and s.e.m. **d**, Cell size distributions measured using a Coulter counter for the indicated strains. Cells were grown on synthetic complete 2% glycerol 1% ethanol.



Extended Data Figure 5 | Whi5 and Cln3-11A stability. **a**, Whi5-mCherry was expressed from a hormone-inducible promoter³⁴ (*LexApr-WHI5-mCherry*), which was inactivated before the experiment (see Fig. 4d, e). The mean amount of Whi5-mCherry in G1-phase daughter cells was measured for each cell relative to its amount at $t = 0$. The distance between the black arrows indicates the s.e.m. We estimated a half-life > 6 h for cells grown in synthetic complete 2% glucose. **b**, **c**, mCitrine-Cln3-11A was expressed from a *MET25* promoter for **(b)** 1 h on SC-Met 2% glucose or **(c)** > 4.5 h on SC-Met 2% glycerol 1% ethanol. Next, transcription was inactivated by switching cells to media composed of either **(b)** synthetic complete + $10\times$ methionine 2% glucose or **(c)** synthetic complete + $10\times$ methionine 2% glycerol 1% ethanol. So that the cells had sufficient time to inactivate protein synthesis, we began our protein half-life measurement 21 min (33 min for synthetic complete 2% glycerol 1% ethanol) after methionine addition. Data and exponential fit shown for daughter cells in G1 phase. Black line indicates means; grey area indicates s.e.m. The short half-life of Cln3-11A relative to the doubling time of cell

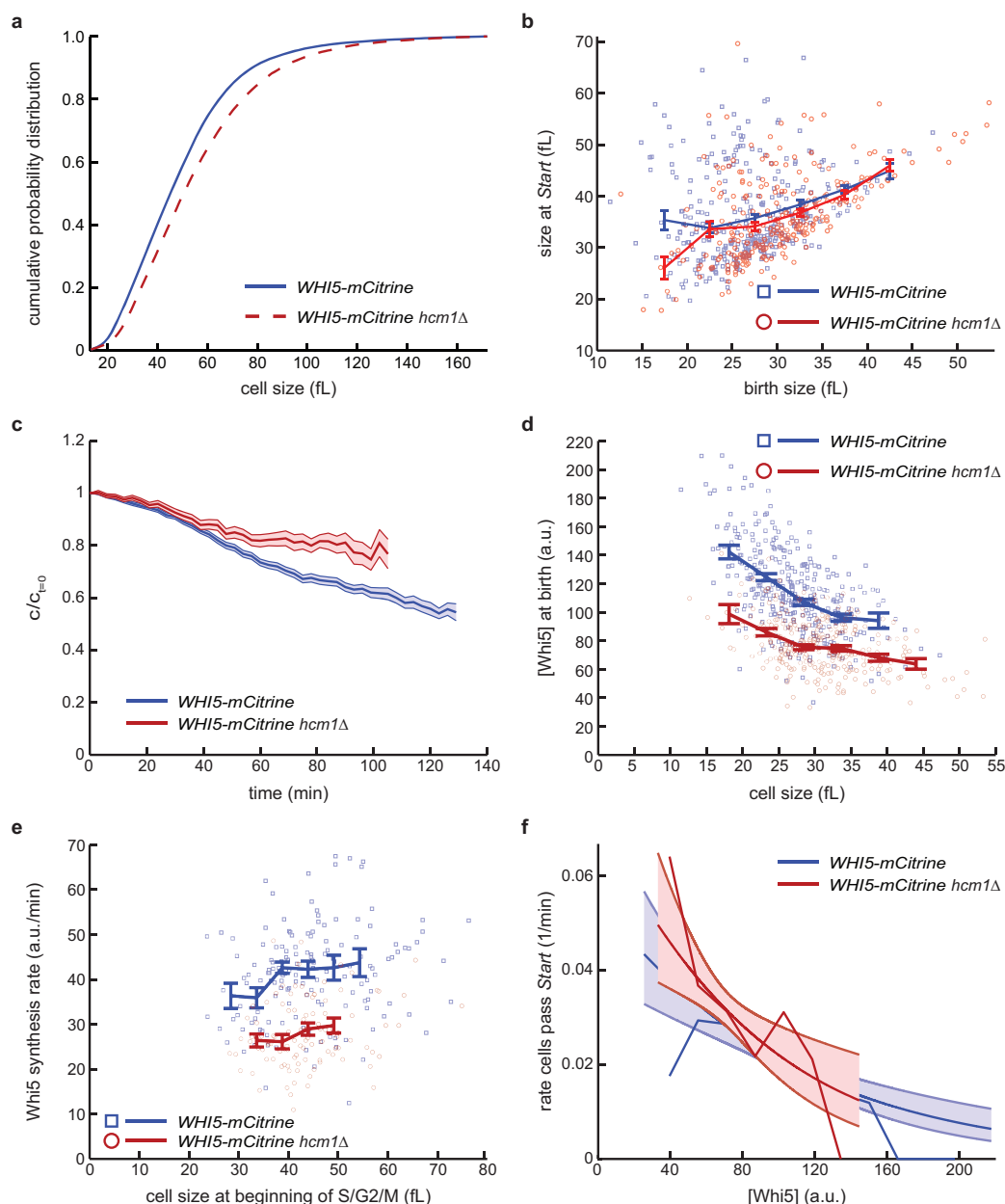
volume, together with the constant concentration of Cln3-11A through G1 (Fig. 1–2), implies that Cln3-11A synthesis is proportional to cell volume. To see this, consider the time-dependent equation for changes in Cln3

concentration $\frac{d[\text{Cln3}]}{dt} = \frac{r}{V} - [\text{Cln3}] \times (d + g)$, where r is the rate of Cln3 protein synthesis (units of molecules \times time⁻¹), V is the cell volume, d is the degradation rate of Cln3 (units of time⁻¹), and g is the rate of dilution of Cln3 due to cell growth (units of time⁻¹). Since $[\text{Cln3}]$ is constant, the left hand side = 0. Also, the half-life of Cln3-11A is larger than that of Cln3, but much smaller than the time it takes to double the cell volume (~ 90 min on SCD, ~ 180 min on synthetic complete 2% glycerol 1% ethanol), so that $d \gg g$. Thus, the equation simplifies to $0 = \frac{r}{V} - [\text{Cln3}] \times d$ so that the rate of Cln3 synthesis is proportional to cell volume, $r = V \times d \times [\text{Cln3}]$. This is consistent with our estimates of Cln3-11A synthesis rates shown in Fig. 2g.



Extended Data Figure 6 | Linking Whi5 partitioning and synthesis to concentration at birth. **a, b**, Daughter cells begin G1 with 1.49 ± 0.03 -fold higher concentration of Whi5 than their mother cells. Shown is the ratio of Whi5–mCitrine concentrations (**a**) and amount (**b**) for daughter–mother pairs at the beginning of G1 phase. **c**, The duration of S/G2/M exhibits small, but significant size-dependence ($P < 0.01$). **d**, The total Whi5 concentration in

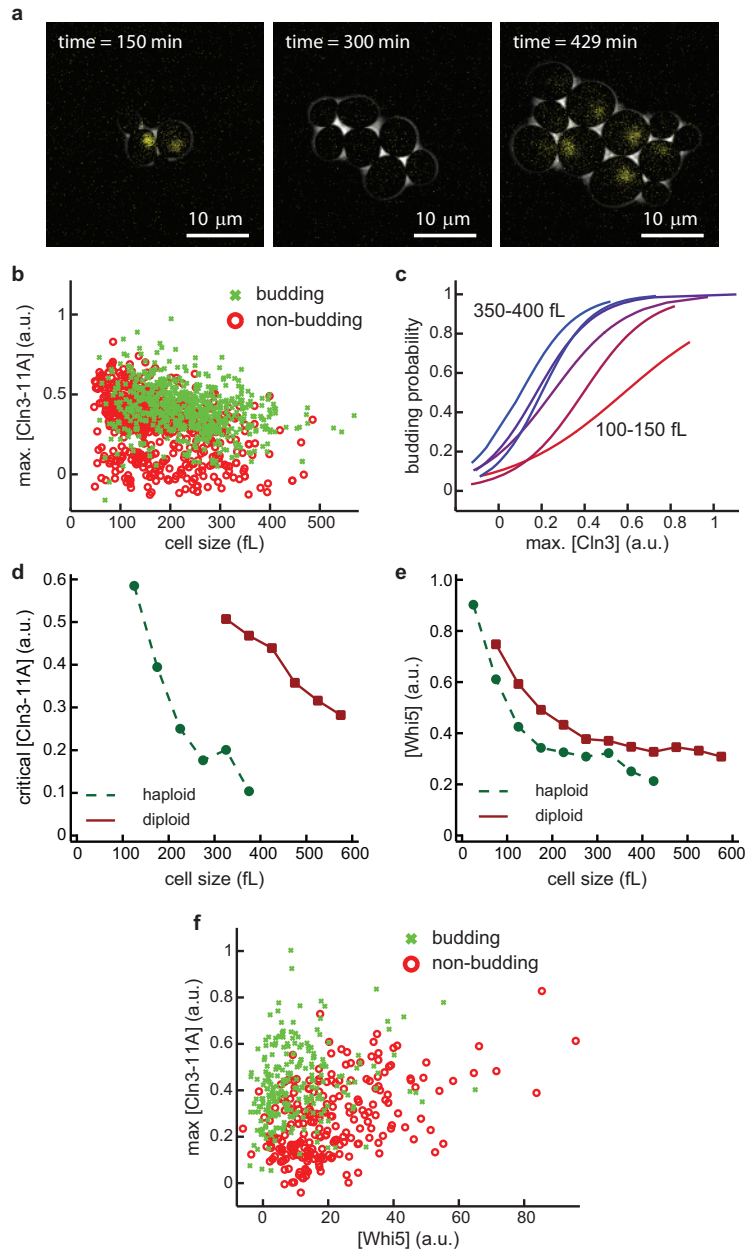
first-generation mother cells just before cytokinesis decreases as a function of cell size. **e**, The size of daughter cells is correlated with the size of their mothers at the time of bud emergence. Cells were grown on synthetic complete 2% glycerol 1% ethanol; $n = 151$. Points denote single-cell data. Bars denote mean values and s.e.m.



Extended Data Figure 7 | Size control and Whi5 synthesis in *hcm1Δ* cells.

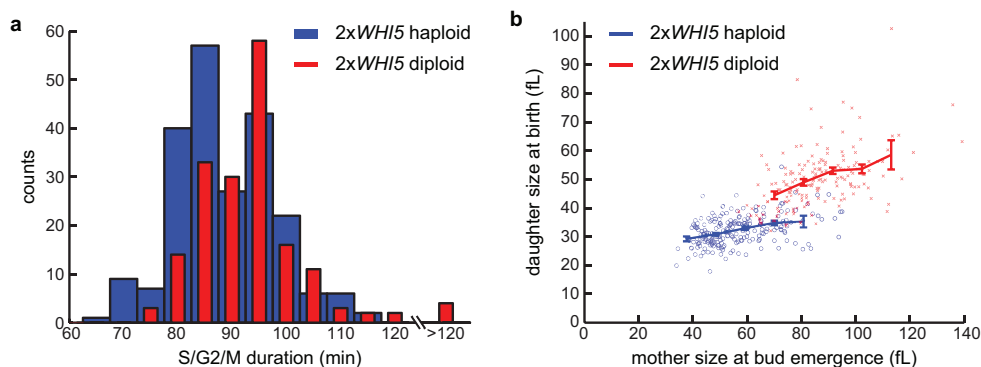
a, Cell size distributions of WT (blue solid line) and *hcm1Δ* (red dashed line) cells, both carrying a *WHI5-mCitrine* allele, were measured in a Coulter counter. **b**, Size at Start as a function of birth size is shown for WT ($n = 339$) and *hcm1Δ* ($n = 262$) daughter cells. Bars denote mean and s.e.m. Note that small *hcm1Δ* cells exhibit poor size control (leftmost bin). **c**, Change in cellular Whi5 concentration during G1 for daughter cells. Cells are born at $t = 0$ and the change in concentration is shown with s.e.m. Blue denotes WT (see also Fig. 1f), red denotes *hcm1Δ* cells. **d**, Whi5 concentration at cell birth is shown as a

function of cell size for WT ($n = 339$) and *hcm1Δ* ($n = 284$) daughter cells. **e**, The rate of Whi5 synthesis as a function of cell size is shown for WT ($n = 151$) and *hcm1Δ* ($n = 106$) cells. Bars denote mean values and s.e.m. Squares and circles denote single-cell data. **f**, The rate at which daughter cells progress through Start is shown as a function of Whi5-mCitrine concentration for WT (blue, $n = 334$) and *hcm1Δ* (red, $n = 262$) cells. Smooth lines are logistic regressions and the corresponding shaded areas denote 95% confidence intervals. Jagged lines connect means for binned data. Cells were grown on synthetic complete 2% glycerol 1% ethanol.



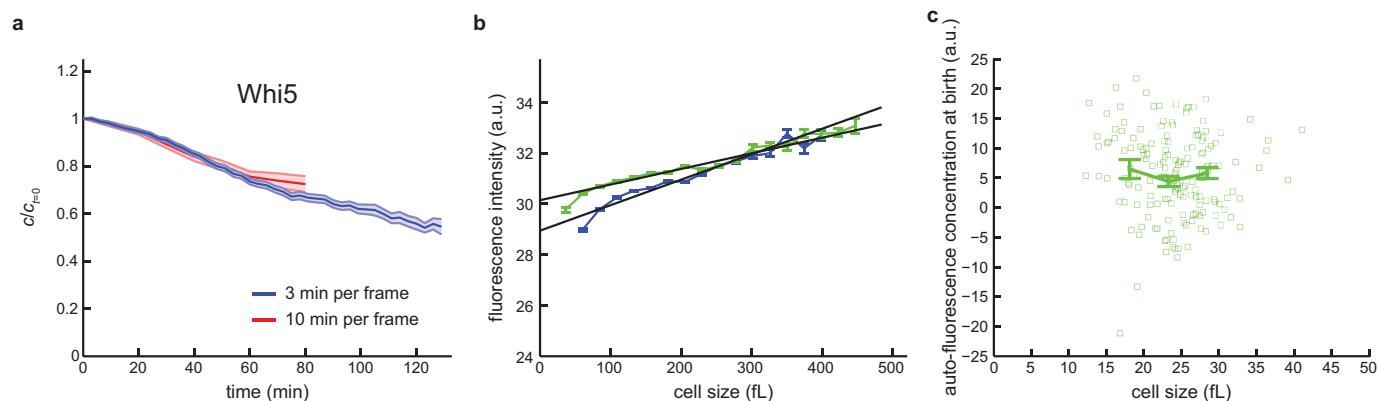
Extended Data Figure 8 | Data supporting Cln3-11A-pulse experiments shown in Fig. 4. a, Composite phase and fluorescence images of *bck2Δ MET25pr-mCitrine-CLN3-11A* haploid cells used in the pulse experiment shown in Fig. 4a–c. Cells were grown in the absence of methionine (*MET25pr-mCitrine-CLN3-11A* on). After 150 min (see image), cells were arrested in G1 by addition of 10× methionine to the SCD-Met medium. After variable lengths of arrest (3 h for the images shown here: see second image), a pulse of Cln3-11A was expressed by removal of methionine from the medium (1 h pulse for the experiment shown here: see third image). **b,** For daughter cells born during the experiment, we determined the maximum Cln3-11A concentration during the pulse, the corresponding cell volume, and whether the cell budded. Data from 22 different experiments were pooled. **c,** Cells were binned according to their size. For each 50 fl size bin, we used a logistic regression to calculate budding probability as a function of Cln3-11A peak concentration. **d,** For each size bin, the critical Cln3-11A concentration was determined as the amplitude of the pulse where 50% of the cells budded. A similar set of experiments was done for diploid cells; $n = 1195$ for haploids, and $n = 405$ for diploids. **e,** The *bck2Δ MET25pr-CLN3-11A WHI5-mCitrine* cells were arrested in G1 by addition of 10× methionine to the SCD-Met medium. Cells were tracked

during the G1 arrest and the median Whi5-mCitrine concentration was measured as a function of size; $n = 162$ for haploids, and $n = 148$ for diploids. **f,** Single-cell data corresponding to Fig. 4e. *MET25pr-mCitrine-CLN3-11A LexApr-WHI5-mCherry bck2Δ* haploid cells were used for Cln3-11A pulse experiments to decouple cell size and Whi5 concentration. Maximum Cln3-11A concentration, corresponding cell size and Whi5 concentration, and whether or not the cell budded were determined in 18 independent experiments for a total of 471 daughter cells (see Methods). This generated a four-dimensional data set that we used to build a logistic regression model. In this model, we predicted cell cycle entry (budding) using a linear combination of cell size, Whi5, and Cln3-11A. This resulted in a model based solely on Cln3-11A and Whi5. Thus, once Cln3-11A and Whi5 concentrations are measured, cell size yields no additional information. To visualize this result in Fig. 4e, we binned our data into six bins based on cell size (greater or less than 295 fl) and Whi5 concentration (<10, 10–25, and 25–40 arbitrary units). For each of these six bins, we performed a logistic regression to estimate the probability of entering the cell cycle as a function of the peak mCitrine-CLN3-11A concentration produced by the pulse.



Extended Data Figure 9 | Ploidy increases S/G2/M duration and cell size at birth. **a**, Histogram showing the duration of S/G2/M for haploid cells containing an extra copy of *WHI5* (blue, $n = 220$) and WT diploid cells (red, $n = 176$). **b**, The size of daughter cells is shown as a function of the size of their

mothers at the time of bud emergence. At a given mother size, diploid cells produce larger daughter cells. Cells were grown on synthetic complete 2% glycerol 1% ethanol. Bars denote means and s.e.m.



Extended Data Figure 10 | Photobleaching control and size-dependent background subtraction. **a**, The concentration of Whi5-mCitrine decreases during G1, as shown in Fig. 1f. Increasing the time between frames from 3 min ($n = 339$) to 10 min ($n = 75$) did not significantly affect our concentration measurements, indicating that photobleaching of the mCitrine fluorescent protein was not significant in our experiments. **b**, Auto-fluorescent signal in the mCitrine channel during G1 arrest for an unlabelled strain (*bck2Δ MET25pr-CLN3*) in two independent experiments (blue: $n = 79$; green: $n = 89$). Cells were grown on SCD + 10× methionine. Bars denote mean and s.e.m. for each size bin. The average of these two experiments was used for background subtraction in Fig. 4. A similar size-dependent background subtraction was performed for each experimental condition and for the mCherry red fluorescent channel. **c**, Cell-to-cell variation in background-subtracted auto-

fluorescence concentration measured in an unlabelled cell. One of the experiments with the unlabelled WT strain used to determine the auto-fluorescence signal for the experiments shown in Fig. 1–3 is analysed the same way as experiments shown in Fig. 2a, b ($n = 164$). This illustrates cell-to-cell variation in auto-fluorescence. Owing to experiment-to-experiment variation, a single control experiment with an unlabelled strain will typically result in a mean ‘concentration’ of ± 5 arbitrary units (compared with the average autofluorescence used for analysis), while cell-to-cell variability in autofluorescence within one experiment exhibits a standard deviation of ~ 10 arbitrary units. Note that the arbitrary units in **a** and **b** are not comparable, because different settings were used to export the microscopy data for the pulse experiments shown in Fig. 4.

Mediator kinase inhibition further activates super-enhancer-associated genes in AML

Henry E. Pelish^{1*}, Brian B. Liao^{1*}, Ioana I. Nitulescu¹, Anupong Tangpeerachaikul¹, Zachary C. Poss², Diogo H. Da Silva¹, Brittany T. Caruso¹, Alexander Arefolov¹, Olugbeminiyi Fadeyi¹, Amanda L. Christie³, Karrie Du¹, Deepti Banka⁴, Elisabeth V. Schneider^{5,6}, Anja Jestel⁵, Ge Zou¹, Chong Si¹, Christopher C. Ebmeier², Roderick T. Bronson⁷, Andrei V. Krivtsov⁸, Andrew G. Myers¹, Nancy E. Kohl³, Andrew L. Kung⁹, Scott A. Armstrong⁸, Madeleine E. Lemieux¹⁰, Dylan J. Taatjes² & Matthew D. Shair¹

Super-enhancers (SEs), which are composed of large clusters of enhancers densely loaded with the Mediator complex, transcription factors and chromatin regulators, drive high expression of genes implicated in cell identity and disease, such as lineage-controlling transcription factors and oncogenes^{1,2}. BRD4 and CDK7 are positive regulators of SE-mediated transcription^{3–5}. By contrast, negative regulators of SE-associated genes have not been well described. Here we show that the Mediator-associated kinases cyclin-dependent kinase 8 (CDK8) and CDK19 restrain increased activation of key SE-associated genes in acute myeloid leukaemia (AML) cells. We report that the natural product cortistatin A (CA) selectively inhibits Mediator kinases, has anti-leukaemic activity *in vitro* and *in vivo*, and disproportionately induces upregulation of SE-associated genes in CA-sensitive AML cell lines but not in CA-insensitive cell lines. In AML cells, CA upregulated SE-associated genes with tumour suppressor and lineage-controlling functions, including the transcription factors *CEBPA*, *IRF8*, *IRF1* and *ETV6* (refs 6–8). The BRD4 inhibitor I-BET151 downregulated these SE-associated genes, yet also has anti-leukaemic activity. Individually increasing or decreasing the expression of these transcription factors suppressed AML cell growth, providing evidence that leukaemia cells are sensitive to the dosage of SE-associated genes. Our results demonstrate that Mediator kinases can negatively regulate SE-associated gene expression in specific cell types, and can be pharmacologically targeted as a therapeutic approach to AML.

CDK8 associates with CCNC (cyclin C), MED12 and MED13 to form a CDK8 module that can reversibly associate with the 26-subunit Mediator complex⁹. Because SEs are disproportionately loaded with Mediator², we examined whether CDK8, as a Mediator-associated kinase, might regulate SE function. Using chromatin immunoprecipitation followed by sequencing (ChIP-seq), we mapped the genome-wide occupancy of CDK8, along with known SE-associated factors and histone modifications, in the AML cell line MOLM-14. Semi-supervised hierarchical clustering revealed that CDK8 most closely associated with MED1, followed by BRD4 and histone 3 Lys27 acetylation (H3K27ac), at putative enhancer elements marked with H3K4me1 (red bar, Fig. 1a and Extended Data Fig. 1a–c). A fraction of these regions was particularly large and loaded with CDK8, MED1 and BRD4, suggesting that they may represent SEs. Consistent with this notion, most of the CDK8, MED1, BRD4 and H3K27ac ChIP-seq signal was disproportionately located on a small number of SEs identified by each factor separately (Extended Data Fig. 1d–f). These SEs significantly overlapped (Fig. 1b, c and Supplementary Table 1). Genes associated with these SEs were enriched with Gene Ontology (GO)

terms pertinent to haematopoiesis, cellular differentiation and transcription, supporting the notion that SEs regulate cellular identity (Supplementary Table 1).

To determine whether pharmacological inhibition of Mediator kinases regulates SE function and inhibits AML proliferation, in analogy to BRD4, we characterized CA (Fig. 2a) as an inhibitor of CDK8 and its paralogue CDK19 (77% identical overall and 94% in the catalytic domain). CA was reported to bind CDK8 and CDK19, as well as ROCK1 and ROCK2, as individual proteins *in vitro*¹⁰. We synthesized CA^{11,12} and determined that it potently inhibited the kinase activity of the CDK8 module *in vitro* (half-maximum inhibitory concentration (IC₅₀) = 12 nM; Fig. 2b and Extended Data Fig. 2a). By contrast, CA did not inhibit other transcriptional cyclin-dependent kinases CDK7 (TFIIH), CDK9 (P-TEFb), CDK12 or CDK13 *in vitro*, nor did it bind CDK9, CDK12, CDK13, ROCK1 or ROCK2 up to 2,500 nM in

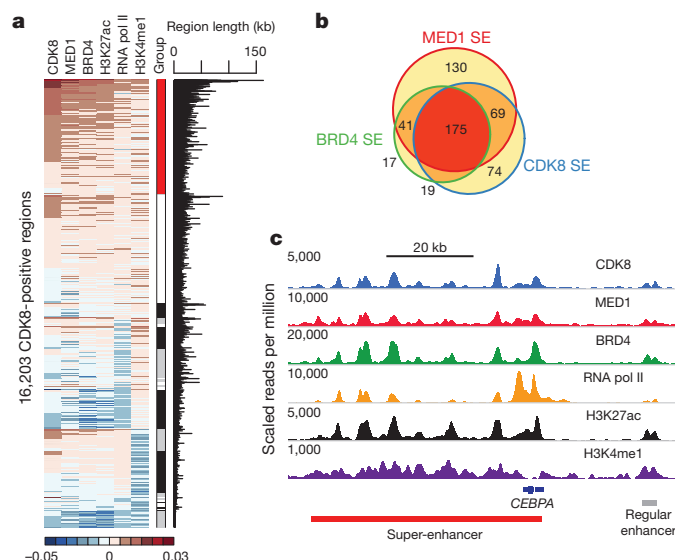


Figure 1 | CDK8 is asymmetrically loaded at SEs in MOLM-14 cells.

a, Clustering of total ChIP-seq signal of CDK8, MED1, BRD4, H3K27ac, RNA pol II and H3K4me1 on CDK8-positive regions. Each respective cluster is ordered by CDK8 signal. The red bar indicates the cluster most highly enriched for the factors listed above. **b**, Overlap between SEs independently identified by ChIP-seq signal for CDK8, MED1 and BRD4 based on the collapsed superset of regions identified by any one factor. **c**, ChIP-seq binding profiles at the *CEBPA* locus.

¹Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, USA. ²Department of Chemistry and Biochemistry, University of Colorado, Campus Box 596, Boulder, Colorado 80303, USA. ³Lurie Family Imaging Center, Dana-Farber Cancer Institute, Boston, Massachusetts 02215, USA. ⁴Division of Hematology/Oncology, Children's Hospital, Boston, Massachusetts 02215, USA. ⁵Proteros Biostructures GmbH, Bunsenstrasse 7a, D-82152 Martinsried, Germany. ⁶Max-Planck-Institut für Biochemie, Am Klopperspitz 18, D-82152 Martinsried, Germany. ⁷Dana-Farber Cancer Institute, Boston, Massachusetts 02215, USA. ⁸Cancer Biology and Genetics Program and Department of Pediatrics, Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA. ⁹Department of Pediatrics, Columbia University Medical Center, New York, New York 10032, USA. ¹⁰Bioinfo, Plantagenet, Ontario K0B 1L0, Canada.

*These authors contributed equally to this work.

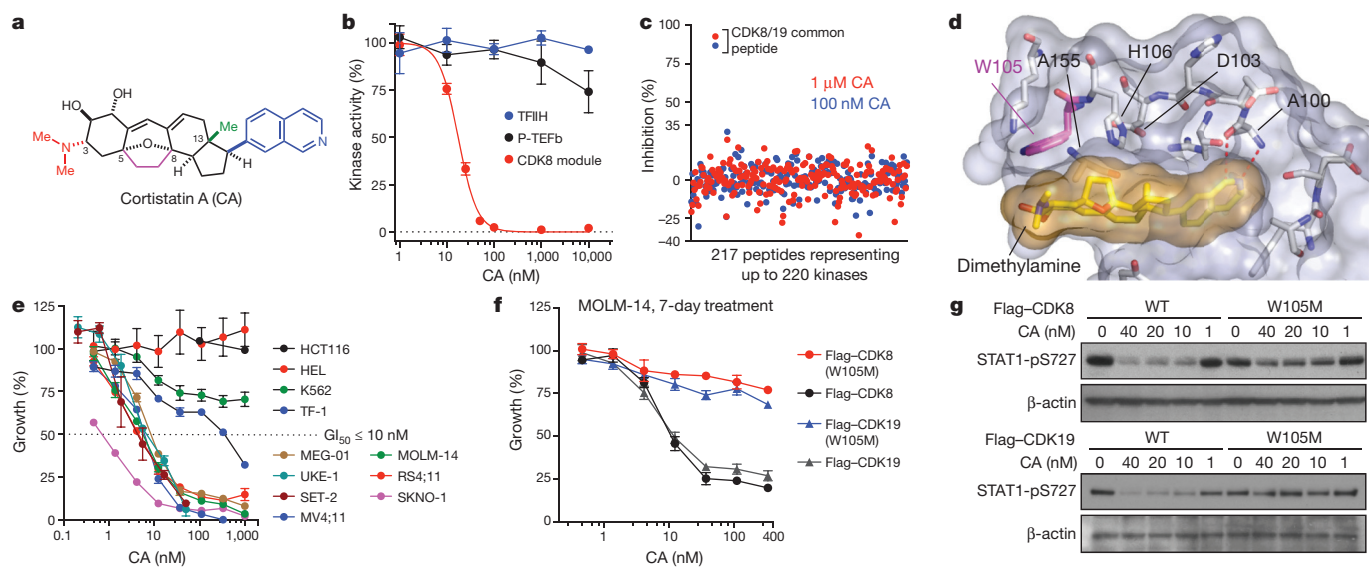


Figure 2 | CA suppresses AML cell proliferation by inhibiting Mediator kinases. **a**, CA structure with N,N-dimethylamine in red, C5–C8 ethano bridge in magenta, C13-methyl in green and isoquinoline in blue. **b**, Phosphorylation of the RNA pol II CTD (mean \pm s.e.m., $n = 3$ biological replicates, one of two experiments shown, autorad in Supplementary Fig. 1). **c**, Kinome profiling in MOLM-14 lysate (mean, $n = 2$ biological replicates, experiment performed once, values $< 35\%$ indicate no change). **d**, CA-binding pocket of CDK8 from CA-CDK8-CCNC crystal structure (semi-transparent surface; CA in

MOLM-14 cell lysate (Fig. 2b and Extended Data Fig. 2b, c). In cells, CA dose-dependently inhibited the phosphorylation of known CDK8 substrates STAT1-S727 (ref. 13; $IC_{50} < 10$ nM), Smad2-T220 and Smad3-T179 (ref. 14; $IC_{50} < 100$ nM) (Extended Data Fig. 2d). No kinase substrates have been reported for CDK19.

We more broadly evaluated CA selectivity in cell lysate (using KiNativ, see Methods)¹⁵ and *in vitro*, which collectively tested 387 kinases. At 100-times its CDK8 IC_{50} value, CA was fully selective in MOLM-14 cell lysate for CDK8 and CDK19, and *in vitro* only inhibited the CDK8-CCNC complex and GSG2, the latter of which we disqualified as a cellular target of CA (Fig. 2c, Extended Data Fig. 2c, e–h, Supplementary Table 2 and Supplementary Information). CA also exhibited high affinity binding (equilibrium dissociation constant (K_d) = 195 ± 15.8 pM (mean \pm s.e.m.)), slow binding kinetics (dissociation rate constant (k_{off}) = $6.35 \times 10^{-5} \pm 8.15 \times 10^{-6}$ s $^{-1}$, association rate constant (k_{on}) = $3.26 \times 10^5 \pm 1.54 \times 10^4$ s $^{-1}$ M $^{-1}$) and a long residence time (262 ± 34 min) in its interaction with CDK8-CCNC complex *in vitro*.

To understand how CA inhibits CDK8, we obtained a high-resolution (2.4 Å) crystal structure of a CA-CDK8-CCNC ternary complex (Fig. 2d and Extended Data Fig. 3). CA exhibits exquisite shape complementarity with the ATP-binding pocket of CDK8. In particular, the isoquinoline of CA forms N–H and CH–O hydrogen bonds with Ala100 (ref. 16), the C5–C8 ethano bridge and the C13-methyl group of CA occupy deep hydrophobic crevices in the ATP-binding site, and the protonated C3 N,N-dimethylamine of CA engages in an apparent cation– π interaction with Trp105 (ref. 17).

We investigated the antiproliferative activity of CA and observed that it inhibited the proliferation (half-maximum growth inhibition concentration (GI_{50}) < 10 nM) of several myeloid, mixed-lineage and megakaryoblastic leukaemia cell lines containing diverse oncogenic contributors, including mixed lineage leukaemia (MLL) fusions (MOLM-14, MV4;11 and RS4;11 cells), *RUNX1-RUNX1T1* (SKNO-1), *JAK2(V617F)* (SET-2 and UKE-1) and *BCR-ABL* (MEG-01) (Fig. 2e, Extended Data Table 1 and Extended Data Fig. 4a). CA inhibited CDK8 kinase activity in both sensitive and insensitive cell lines with similar potency, and did not alter CDK8 or CDK19 protein levels

(Extended Data Fig. 4b, c). Although SET-2 and HEL cell lines contain the *JAK2(V617F)* mutation, and MEG-01 and K562 contain the *BCR-ABL* translocation, megakaryoblastic cell lines SET-2 and MEG-01 cells were sensitive to CA whereas erythroleukaemia-derived cell lines HEL and K562 were not, suggesting that cell lineage may be a contributing determinant for CA sensitivity¹⁸. The phenotypic effects of CA were cell-line-dependent. CA treatment increased megakaryocyte markers CD41 and CD61 on SET-2 cells, whereas CA treatment of MOLM-14, MV4;11 and SKNO-1 cells increased cleaved PARP levels, annexin V staining and the sub-G1 cell population, consistent with apoptosis (Extended Data Fig. 4d–f).

We confirmed that Mediator kinases mediate the antiproliferative activity of CA by identifying a point mutant of CDK8 and CDK19, W105M, that maintained catalytic activity but specifically conferred resistance to CA (Fig. 2f, g and Extended Data Fig. 5a–f). Notably, CDK8 and CDK19 are the only mammalian cyclin-dependent kinases with Trp (or any aromatic amino acid) at residue 105 (Extended Data Fig. 5g), underscoring the importance of the putative cation– π interaction.

Next, we used CA to investigate whether Mediator kinase activity regulates SE-associated gene expression in AML cells. Global gene expression profiling in MOLM-14 cells treated with CA revealed that genes upregulated by CA at 3 h were highly enriched for association with SEs by gene set enrichment analysis (GSEA)¹⁹ (Fig. 3a, b, Extended Data Fig. 6a and Supplementary Table 3). These SE-associated gene sets ranked among the most significantly enriched compared to all other signatures tested (Fig. 3c). Genes upregulated (≥ 1.2 -fold) by CA were disproportionately associated with SEs in MOLM-14 cells (49 out of 251, 20%) compared to regular enhancers (173 out of 5,034, 3%) (Extended Data Fig. 6b, Fisher's exact test, $P < 2.2 \times 10^{-16}$). By contrast, of 102 genes downregulated (≥ 1.2 -fold) by CA, only three were identified as SE-associated (3 out of 102, 3%). Furthermore, the association between CA-upregulated genes (≥ 1.2 -fold) and SE-associated genes correlated with CDK8 occupancy (Fisher's exact test, $P = 2.5 \times 10^{-8}$), consistent with the notion that SEs are direct targets of CA treatment in MOLM-14 cells (Extended Data Fig. 6b).

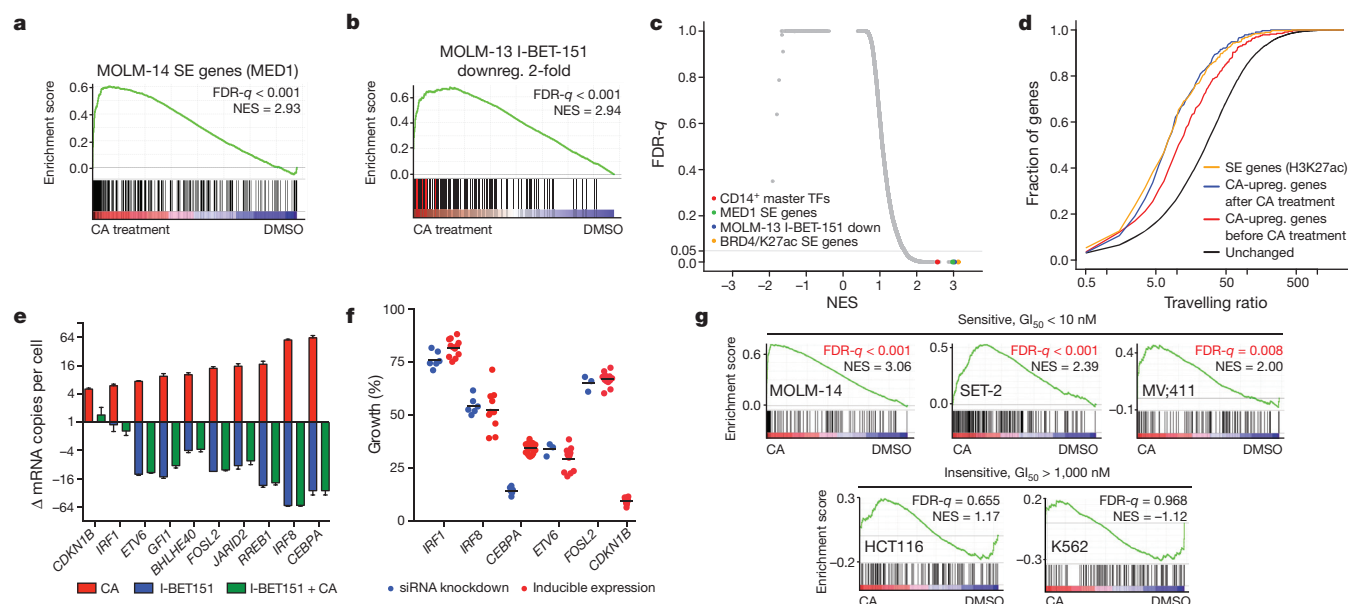


Figure 3 | CA disproportionately increases transcription of SE-associated genes. **a**, **b**, GSEA plots show that genes upregulated after 3h CA treatment of MOLM-14 cells are significantly enriched in MOLM-14 SE-associated genes (**a**), and genes downregulated by IBET-151 ≥ 2 -fold in MOLM-13 cells (**b**). Red bars in **b** indicate H3K27ac SE genes in MOLM-14 cells in GSEA leading edge (22 genes, Fisher's exact test, $P = 1.2 \times 10^{-3}$). **c**, Scatterplot of false discovery rate (FDR- q) versus normalized enrichment score (NES) for indicated gene sets evaluated by GSEA ($n = 3,867$), including C2 of MSigDB. **d**, Cumulative distribution plots of RNA pol II travelling ratio. **e**, Change in

mRNA copy number per cell of selected SE genes after 3 h treatment (red and blue bars) or after 6 h I-BET151 treatment with CA treatment for the final 3 h (green bar) (mean \pm s.e.m., $n = 3$ biological replicates, one of two experiments shown). **f**, Effect of change in expression of selected SE genes on MOLM-14 cell growth (mean \pm s.e.m., with $n = 3$ biological replicates for siETV6 and siFOSL2 and 6 for other siRNA knockdowns, 24 for Flag-CEBPA and 12 for other inducible expressions, one of 2–6 experiments shown). **g**, GSEA of SE genes in CA-treated cells. Regions of CDK8 and H3K27ac co-enrichment identify SE genes in each cell line.

Because SE-associated genes are more highly expressed compared to regular enhancer-associated genes, we determined whether genes upregulated by CA had elongating RNA polymerase (pol) II and reduced travelling ratios²⁰ (ratio of RNA pol II ChIP-seq reads in the proximal promoter versus the gene body). Indeed, CA-upregulated genes exhibited a reduced baseline travelling ratio (2.40-fold, $P < 2.2 \times 10^{-16}$, red versus black curve, Fig. 3d and Extended Data Fig. 6c, d), consistent with CA upregulating active genes, including those associated with SEs. CA treatment further reduced the travelling ratio of these 'CA-upregulated' genes to a level similar to all SE-associated genes (yellow curve), in agreement with their increased expression after CA treatment (1.48-fold, $P = 7.6 \times 10^{-4}$, blue versus red curve, Fig. 3d). Genes downregulated by CA experienced insignificant changes in the travelling ratio (Extended Data Fig. 6e). Global effects of CA on the RNA pol II travelling ratio, RNA pol II carboxy-terminal domain (CTD) phosphorylation, messenger RNA and total RNA levels were modest or negligible (Extended Data Fig. 6f–h).

We then examined whether the upregulation of SE-associated genes might contribute to the antiproliferative activity of CA. SE-associated genes upregulated by CA were enriched in lineage-controlling master transcription factors identified in related CD14⁺ monocytes¹, including tumour suppressors *IRF1*, *IRF8*, *CEBPA* and *ETV6* (Fig. 3e and Extended Data Fig. 7a–c). Increased expression of these genes individually, as well as SE-associated genes *FOSL2* and *CDKN1B*, inhibited the proliferation of MOLM-14 cells (Fig. 3f and Extended Data Fig. 7d, e). ChIP-seq data revealed CDK8 occupancy at the nearby SE of each gene (*CEBPA*, Fig. 1c; and *ETV6* and *FOSL2*, Extended Data Fig. 7f). Furthermore, expression of CA-resistant CDK8(W105M) prevented upregulation of SE-associated genes by CA (Extended Data Fig. 7g). Therefore, upregulation of SE-associated genes, through Mediator kinase inhibition, could contribute to the antiproliferative activity of CA.

Growth of several AML cell lines was sensitive to CA and the BRD4 inhibitor I-BET151 (Extended Data Table 1). The opposing effects of these inhibitors on SE-associated gene expression (Fig. 3b, e (red ticks are

SE-associated genes) and Extended Data Fig. 7c), however, suggest that AML cells might depend on a precise 'dosage' of SE-associated gene expression. Indeed, MOLM-14 cell growth was inhibited by either reduced or increased expression of the same SE-associated genes, many of which were upregulated by CA and downregulated by I-BET151 (Fig. 3e, f and Extended Data Fig. 7c–e, h). Despite having opposing effects on SE-associated genes, CA and I-BET151 co-treatment did not normalize transcription of these genes. Instead, I-BET151-induced transcriptional effects dominated, suggesting a dependence on BRD4 for CA-induced transcription (Fig. 3e and Extended Data Fig. 7c). Consistent with this, I-BET151 caused reduced occupancy of BRD4 and CDK8 on enhancer regions, and CA and I-BET151 co-treatment inhibited MOLM-14 cell growth (Extended Data Fig. 7i, j).

We extended our gene expression, ChIP-seq and SE analyses to additional cell lines that were sensitive (SET-2 and MV4;11) and insensitive (HCT116 and K562) to CA, and found that only the sensitive cell lines showed statistically significant enrichment of SE-associated genes among those upregulated by CA (Fig. 3g). These results support upregulation of SE-associated genes as contributing to the antiproliferative effects of CA. However, we cannot exclude the contribution of other factors.

Finally, we assessed the *in vivo* antileukaemic activity of CA. We first determined that CA had acceptable pharmacokinetic properties in mice for once-daily intraperitoneal dosing (Extended Data Fig. 8a) and then measured its efficacy in a disseminated human AML model²¹. CA afforded a dose-dependent reduction in disease progression ($P < 0.0001$), spleen weight, leukaemia cell burden, and survival (29.5-day median extension in survival, $P < 0.0001$; Fig. 4a, b and Extended Data Fig. 8b–e). Efficacious dosing was well-tolerated, with no loss in body weight or deleterious effects in peripheral blood of leukaemia-bearing or healthy, immunocompetent (CD-1) mice (Extended Data Fig. 8f, g, j–l). In a second AML model using SET-2 cells, CA afforded a 71% tumour volume reduction, also with no loss in body weight (Fig. 4c and Extended Data Fig. 8h). We confirmed that

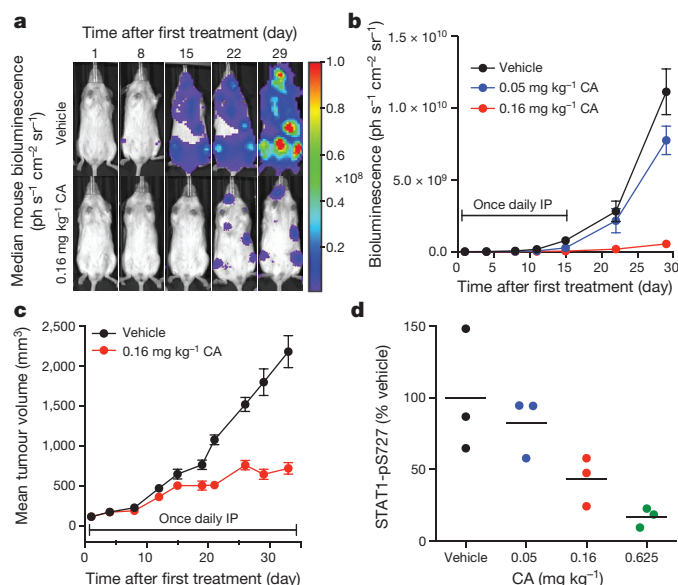


Figure 4 | CA inhibits AML progression and CDK8 *in vivo*.

a, Bioluminescent images of mice bearing MV4;11 leukaemia cells. Mouse with median bioluminescence shown, treatment as in **b**. Colour scale 1.00×10^8 to 1.00×10^6 . **b**, Mean \pm s.e.m., $n = 11$ mice; $P < 0.0001$ for both doses on day 33 versus vehicle, two-way analysis of variance (ANOVA). IP, intraperitoneal. **c**, Mice containing SET-2 AML xenograft tumours and treated as indicated. Mean \pm s.e.m., $n = 10$ mice; 71% tumour growth inhibition on day 33, $P < 0.0001$, two-tailed *t*-test. **d**, Densitometric analysis of STAT1-pS727 in natural killer cells isolated from the spleen of C57BL/6 mice treated with CA or vehicle ($n = 3$ mice), STAT1-pS727 normalized to actin, $P = 0.011$ for 0.625 mg kg^{-1} , one-way ANOVA, experiment performed once.

CA inhibited CDK8 *in vivo* by observing a dose-dependent reduction in STAT1-S727 phosphorylation in natural killer cells, which have CDK8-dependent constitutively phosphorylated STAT1-S727 (Fig. 4d and Extended Data Fig. 8i)²².

Although SE-associated genes are expressed at high levels, our results with CA show that a subset is restrained from even higher expression by Mediator kinase activity. The specificity, potency, favourable pharmacokinetics and long residence time of CA make it a useful *in vitro* and *in vivo* probe of Mediator kinases and a promising lead for development of therapeutics.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 27 December 2014; accepted 14 July 2015.

Published online 28 September 2015.

- Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–947 (2013).
- Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–319 (2013).
- Lovén, J. *et al.* Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell* **153**, 320–334 (2013).
- Dawson, M. A. *et al.* Recurrent mutations, including NPM1c, activate a BRD4-dependent core transcriptional program in acute myeloid leukemia. *Leukemia* **28**, 311–320 (2013).
- Kwiatkowski, N. *et al.* Targeting transcription regulation in cancer with a covalent CDK7 inhibitor. *Nature* **511**, 616–620 (2014).
- Prange, K. H. M., Singh, A. A. & Martens, J. H. A. The genome-wide molecular signature of transcription factors in leukemia. *Exp. Hematol.* **42**, 637–650 (2014).
- Fragale, A., Marsili, G. & Battistini, A. Genetic and epigenetic regulation of interferon regulatory factor expression: implications in human malignancies. *J. Genet. Syndr. Gene Ther.* **4**, 205 (2013).

- de Braekeleer, E. *et al.* ETV6 fusion genes in hematological malignancies: a review. *Leuk. Res.* **36**, 945–961 (2012).
- Allen, B. L. & Taatjes, D. J. The Mediator complex: a central integrator of transcription. *Nature Rev. Mol. Cell Biol.* **16**, 155–166 (2015).
- Cee, V. J., Chen, D. Y.-K., Lee, M. R. & Nicolaou, K. C. Cortistatin A is a high-affinity ligand of protein kinases ROCK, CDK8, and CDK11. *Angew. Chem. Int. Edn Engl.* **48**, 8952–8957 (2009).
- Lee, H. M., Nieto-Oberhuber, C. & Shair, M. D. Enantioselective synthesis of (+)-cortistatin A, a potent and selective inhibitor of endothelial cell proliferation. *J. Am. Chem. Soc.* **130**, 16864–16866 (2008).
- Flyer, A. N., Si, C. & Myers, A. G. Synthesis of cortistatins A, J, K and L. *Nature Chem.* **2**, 886–892 (2010).
- Bancerek, J. *et al.* CDK8 kinase phosphorylates transcription factor STAT1 to selectively regulate the interferon response. *Immunity* **38**, 250–262 (2013).
- Alarcón, C. *et al.* Nuclear CDKs drive Smad transcriptional activation and turnover in BMP and TGF- β pathways. *Cell* **139**, 757–769 (2009).
- Patricelli, M. P. *et al.* *In situ* kinase profiling reveals functionally relevant properties of native kinases. *Chem. Biol.* **18**, 699–710 (2011).
- Pierce, A. C., Sandretto, K. L. & Bernis, G. W. Kinase inhibitors and the case for CH...O hydrogen bonds in protein-ligand binding. *Proteins* **49**, 567–576 (2002).
- Zacharias, N. & Dougherty, D. A. Cation- π interactions in ligand recognition and catalysis. *Trends Pharmacol. Sci.* **23**, 281–287 (2002).
- Garraway, L. A. & Sellers, W. R. Lineage dependency and lineage-survival oncogenes in human cancer. *Nature Rev. Cancer* **6**, 593–602 (2006).
- Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
- Adelman, K. & Lis, J. T. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nature Rev. Genet.* **13**, 720–731 (2012).
- Etchin, J. *et al.* Antileukemic activity of nuclear export inhibitors that spare normal hematopoietic cells. *Leukemia* **27**, 66–74 (2013).
- Putz, E. M. *et al.* CDK8-Mediated STAT1-S727 phosphorylation restrains NK cell cytotoxicity and tumor surveillance. *Cell Rep.* **4**, 437–444 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank R. Levine, R. King, B. Ebert, B. Bernstein, S. Gillespie, M. Galbraith, M. Patricelli and T. Nomanbhoy for discussions. Lentiviral packaging was completed at the University of Massachusetts Medical School RNAi core facility. Microarray data collection was performed at DFCI MicroArray Core Facility and UMass Medical School Genomics Core Facility. Formulation was performed at VivoPath. *In-vivo* portions of pharmacokinetic, natural killer and SET-2 studies were performed at Charles River. We thank S. Trauger and G. Byrd of Harvard FAS Small Molecule Mass Spectrometry for PK data acquisition and Harvard FAS Center for Systems Biology for flow sorting and high-throughput sequencing. Recombinant expression of CDK8 module subunits was completed at the Tissue Culture Shared Resource at the University of Colorado Cancer Center, supported by the NCI (P30 CA046934). HCT116 RNA-seq was carried out at the Genomics Shared Resource at the University of Colorado Cancer Center and supported by grant P30-CA046934. We thank A. Odell and R. Dowell for HCT116 RNA-seq data analysis, the R. Levine laboratory (MSKCC) for carrying out the SET-2 RNA-seq acquisition, the M. Geyer laboratory for purified CDK12–CCNK and CDK13–CCNK complexes, and P. Kovarik for STAT1 plasmids. This work was supported by NIH grant CA66996 (S.A.A.), NCI grants R01 CA170741 (D.J.T.) and F31 CA180419 (Z.C.P.), NIH T32 GM08759 (Z.C.P.), a Leukemia and Lymphoma Society Translational Research Program Grant (M.D.S.), the Blavatnik Biomedical Accelerator Program at Harvard (M.D.S.) and the Starr Cancer Consortium (M.D.S.).

Author Contributions H.E.P., B.B.L. and M.D.S. designed the research and analysed data. H.E.P., B.B.L., I.J.N., A.T., D.H.D., B.T.C. and K.D. performed cell-based and biochemical experiments not otherwise specified, and analysed data under guidance from M.D.S. Z.C.P. and C.C.E. performed *in vitro* kinase assays and HCT116 gene expression under guidance from D.J.T. A.A. and O.F. synthesized CA under guidance from M.D.S. C.S. and G.Z. synthesized CA under guidance from A.G.M. A.L.C. performed MV4;11 *in vivo* efficacy and safety studies under guidance from N.E.K. D.B. performed early MOLM-14 cell growth assays under guidance of S.A.A. E.V.S. and A.J. performed X-ray crystallography. R.T.B. performed mouse histopathology. A.L.K. advised on *in vivo* studies. S.A.A. and A.V.K. advised on AML studies. M.E.L. performed computational biology studies. H.E.P., B.B.L., D.J.T. and M.D.S. wrote the manuscript. M.D.S. supervised the research.

Author Information The atomic coordinates of CDK8–CCNK in complex with cortistatin A have been deposited in the Protein Data Bank (PDB) with accession number 4CRL. MIAME-compliant microarray data as well as aligned and raw ChIP-seq data were deposited to the Gene Expression Omnibus (GEO) with accession GSE65161. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.D.S. (shair@chemistry.harvard.edu).

METHODS

Cell culture. All media was supplemented with 100 U ml⁻¹ penicillin and 100 µg ml⁻¹ streptomycin. Cell line media: MV4;11, RS4;11, K562, HEL, MOLM-14 and MEG-01 in RPMI-1640, 10% FBS; SET-2 in RPMI-1640, 20% FBS; UKE-1 in RPMI-1640, 10% FBS, 10% horse serum and 1 µM hydrocortisone; SKNO-1 and TF-1 in RPMI-1640, 10% FBS, plus 10 and 2 ng ml⁻¹ GM-CSF, respectively; HaCaT in DMEM, 10% FBS; and HCT116 in McCoy's 5A, 10% FBS (proliferation assay) or DMEM, 10% FBS (gene expression study). Sources: HepG2, MV4;11, RS4;11, MEG-01, TF-1, HCT116 and K562 from ATCC; SKNO-1 from DSMZ; HEL, UKE-1 and SET-2 from R. Levine; and HaCaT, MV4;11-mCLP and MOLM-14 from V. Wilson, A. Kung and S. Armstrong, respectively. MOLM-14 cells were authenticated by STR profiling and flow cytometry. All cell lines were routinely tested for mycoplasma.

Reagents. Compounds were stored under argon at -80 °C in 100% DMSO. Vehicle represents 0.1% DMSO unless otherwise specified. Sources: IFN-γ (PHC4031, Life Technologies), TGF-β1 (R&D Systems), paclitaxel (LC Laboratories), I-BET151 (Tocris), PMA (Calbiochem), and doxorubicin and puromycin (Sigma-Aldrich). Immunoblot antibodies: anti-Flag (F1804), anti-actin (A5060) and anti-CDK19 (HPA007053) from Sigma-Aldrich; anti-Smad2/3 (8685), anti-Smad2 pTail (3108), anti-STAT1 (9172), anti-phospho-STAT1 Tyr701 (9170) and anti-phospho-STAT1 Ser727 (9177), anti-CEBPA (2843), anti-ROCK1 (4035), anti-ROCK2 (9029), anti-CDK8 (4101), anti-caspase-3 (9662) anti-PARP (9532) and anti-CDK9 (2316) from Cell Signaling Technology (CST); anti-phospho-Smad2/3 T220/T179 (600-401-C48) from Rockland; anti-CDK12 (NB100-87012) and anti-CDK13 (NB100-68268) from Novus; and anti-CDK8 (A302-501A) and anti-Haspin (A302-241A) from Bethyl. ChIP antibodies: RNA pol II (Rpb1 N terminus, sc-899X lot B2713) from Santa Cruz; MED1 (A300-793A lot A300-793A-2), BRD4 (A301-985A lot A301-985A50-3), and CDK8 (A302-500A lot A302-500A-1) from Bethyl; and H3K4me3 (ab8580 lot 1308511), H3K27ac (ab4729 lot GR104852-1), and H3K4me1 (ab8895 lot GR61306-1) from Abcam.

Kinase assays. Data were quantified with ImageJ and plotted and fitted with GraphPad Prism 6.0. For STAT1 transactivation domain (TAD), 750 ng of glutathione S-transferase (GST)-STAT1 TAD (residues 639–750) was incubated with ~50 ng recombinant CDK8 module at 30 °C for 8 min in kinase buffer (25 mM Tris, pH 8, 2 mM dithiothreitol (DTT), 100 µM cold ATP, 100 mM KCl, 10 mM MgCl₂ and 2.5 µCi [γ -³²P]ATP (Perkin Elmer) per reaction). The assay included 2.5% DMSO, which did not inhibit kinase activity. 12% SDS-PAGE gels were subsequently silver-stained, exposed for 18 h on a Phosphor Screen and imaged (Typhoon 9400, GE Life Sciences). For pol II CTD, 400 ng of GST-CTD (mouse sequence) was incubated with ~40 ng recombinant CDK8 module, 25 ng TFIIF, or 40 ng P-TEFb at 30 °C for 60 min in kinase buffer. Kinase amounts were chosen to give similar total pol II CTD signal. 9% SDS-PAGE gels were silver stained and exposed as above. *In vitro* Flag-CDK8 kinase assays used ~40 ng kinase and 500 ng GST-CTD. CDK12(714–1063)-CCNK(1–267) and CDK13(694–1039)-CCNK(1–267) were expressed in insect cells and used at ~500 nM per reaction. These regions of CDK12/13 encompass the kinase domains (including the C-terminal extension helix) and the cyclin boxes, and are fully phosphorylated in the T-loop. For STAT1 or Smad2/3, cells were treated with compound for 1 h followed by IFN-γ or TGF-β1 for 1 h, then washed twice with cold PBS, and lysed (RIPA buffer with inhibitors R0278, P8340, P0044 and P5762; Sigma-Aldrich). Standard immunoblotting followed. All experiments were performed twice.

Protein purification. Buffers for purification and elution of recombinant proteins included 0.25 mM PMSF, 1 mM DTT, 1 mM benzamidine and 1 mM sodium metabisulphite. TFIIF was captured from HeLa nuclear extract using a monoclonal antibody for the p89 subunit immobilized to Protein A Sepharose (GE). Final purification of peptide-eluted TFIIF was performed on a 1 ml HiTrap Heparin HP (GE) resulting in 0.1–0.2 µM TFIIF. P-TEFb was purified as described²³ with a Superdex 200 polishing resulting in ~0.5 µM P-TEFb. Recombinant CDK8 module was purified as described²⁴ with omission of the glycerol gradient. STAT1 TAD and pol II CTD were expressed as N-terminal GST fusion proteins in *Escherichia coli* BL21-CodonPlus cells to A_{600 nm} 0.5, then induced with 0.5 mM IPTG for 4 h at 30 °C and batch affinity purified with glutathione Sepharose 4B (GE). Cells were lysed in H/E buffer (50 mM Tris, pH 7.9, 0.5 M NaCl, 0.5 mM EDTA, 10% glycerol and 0.5% NP-40), immobilized on glutathione Sepharose 4B in H/E buffer for 3 h at 4 °C and washed with ~100 column volumes of high-salt buffer (50 mM Tris, pH 7.9, 1 M NaCl, 0.5 mM EDTA, 0.5% NP-40 and 8 mM CHAPS), 0.5 M HEGN (20 mM HEPES, pH 7.6, 0.5 M KCl, 0.1 mM EDTA, 10% glycerol and 0.02% NP-40) and 0.15 M HEGN (20 mM HEPES, pH 7.6, 0.15 M KCl, 0.1 mM EDTA, 10% glycerol and 0.02% NP-40). Fusion proteins were eluted in 2× column volumes of 30 mM reduced L-glutathione in GSH elution buffer (80 mM Tris, pH 7.9, 0.15 M KCl,

0.1 mM EDTA, 10% glycerol and 0.02% NP-40). The GST-pol II-CTD was further purified by Superdex 200 polishing. Flag-CDK8 wild-type and W105M mutants were expressed in MOLM-14 cells, captured using anti-Flag M2 affinity resin (Sigma-Aldrich), and eluted with 1 mg ml⁻¹ Flag peptide in 0.15 M HEGN in 1× column volume twice. Flag peptide elutions were stained with SYPRO Ruby to standardize kinase amounts. Purifications contained cyclin C but not MED12 or MED13 (data not shown).

Native kinase capture immunoblot and native kinome-wide profiling. Experiments were performed as previously described^{15,25}. 5 × 10⁸ MOLM-14 cells were washed twice with 10 ml cold PBS and resuspended in 1 ml cold kinase buffer (20 mM HEPES, pH 7.4, 150 mM NaCl, 0.5% Triton X-100, with inhibitors 11697498001, Roche and P5726, Sigma). Cells were lysed by sonication (2 × 10 s pulses with a 30 s break) and centrifuged (16,000g, 10 min). The supernatant was desalted through a column (732-2010, Biorad) and the eluted lysate was diluted to 5 mg ml⁻¹ with kinase buffer. For each treatment, 475 µl of the lysate was pre-incubated with 10 µl MnCl₂ (1 M) and 5 µl compound to the desired concentration at room temperature for 30 min. Uninhibited kinases were captured with 10 µl ActivX desthiobiotin-ATP probe (0.25 mM; 88311, Pierce) at room temperature for 10 min. Samples were mixed with 500 µl urea (8 M; 818710, Millipore) and 50 µl streptavidin agarose (20359, Thermo) for 60 min at room temperature on a nutator. Beads were washed twice with a 1:1 mixture of kinase buffer and 8 M urea, and collected by centrifugation (1,000g, 1 min). Proteins were eluted from the beads with 100 µl 2 × LDS sample buffer (NP0007, Life) at 95 °C for 10 min. Samples were analysed by standard immunoblotting and horseradish peroxidase detection. Experiment was performed twice. Native kinome profiling was performed with MOLM-14 cell lysate according to the KiNativ Method by ActivX Biosciences. For each peptide quantified, the change in mass spectrometry signal for the treated samples relative to the signal for the control samples was expressed as percentage inhibition. The results correspond to one experiment of duplicates for each CA concentration. The percentage changes in mass spectrometry signal reported are statistically significant (Student's *t*-test score <0.04).

Recombinant kinome-wide selectivity profiling and IC₅₀ determination. A radiometric protein kinase assay was used (PanKinase activity assay; performed by ProKinase GmbH) as described²⁶. IC₅₀ determination for CDK8-CCNC (8.3 nM with 1.0 µM ATP and 1.0 µg/50 µl of substrate RBER-IRStide) was performed as duplicate measurements and IC₅₀ was calculated using Prism 5.04 with sigmoidal response, top fixed at 100% and bottom at 0% with least-squares fitting.

Binding and kinetics. Measurements listed were made using the Proteros reporter displacement assay as previously described²⁷. CDK8-CCNC (0.62 nM) was preincubated with a reporter probe at a concentration equal to its binding affinity (*K_d*) in 20 mM MOPS, pH 7.0, 1 mM DTT and 0.01% Tween20 (final reaction volume 10 µl in black polypropylene U bottom plates, Corning 4514). After transfer of serially diluted CA, probe displacement was monitored for 60 min. *K_d* values were calculated using the Cheng-Prusoff equation from the IC₅₀ values obtained from the percentage displacement values at the last time point measured. Association rate constant was calculated from the decay rate of probe displacement. Dissociation rate constant was determined as the product of *K_d* × association rate constant. Residence time was calculated as 1/*k_{off}*. Error was determined by Gaussian error propagation from the IC₅₀ error. Experiment was performed once.

Crystallization, data collection and refinement. Human CDK8-CCNC was expressed and purified as previously described²⁷. Co-crystals at a protein concentration of 11.3 mg ml⁻¹ with 1 mM CA were obtained in 20% PEG 3350 and 0.20 M sodium formate at 20 °C and shock-frozen with 25% ethylene glycol as cryoprotectant. Diffraction data were collected at the Swiss Light Source (SLS, Villigen, Switzerland), beamline X06SA with a wavelength of 1.00004 Å at 100 K, and processed using XDS and XSCALE²⁸. The structure was solved by molecular replacement²⁹, subsequent model building and refinement (including TLS refinement) was performed with COOT³⁰ and CCP4 (refs 31, 32). The *R_{free}* validation was based on a subset of about 3.4% of the reflections omitted during refinement. Waters were included at stereochemically reasonable sites. Final refinement cycles led to a model with *R_{work}* value 21.7% and *R_{free}* value 26.6%. All main-chain angles of non-glycine residues fall into the conformationally most favoured (93.2%), additionally allowed (6.6%) or generously allowed (0.2%) regions of the Ramachandran plot. Graphical figures were prepared using PyMOL³³. Values in parentheses in Extended Data Table 2 refer to the highest resolution shell.

Cell growth assay. All suspension cells were plated (96-well) in triplicate at 5,000–30,000 cells per well for testing (*n* = 3). Viable cell number was estimated after 3, 7 and 10 days by counting viable cells from one vehicle well, generating a cell dilution series, transferring 20 µl per well in duplicate to a 384-well plate, and performing a linear regression to CellTiter-Glo (Promega) response (SPECTRAmax M3, Molecular Devices). Cells from all wells were also fourfold diluted in media and

transferred in duplicate for CellTiter-Glo measurement. On days 3 and 7, an equal volume for all wells was split-back with fresh media and compound, such that the resulting cell density for the vehicle well matched the initial seeding density. For days 7 and 10, estimated cell number represents the split-adjusted theoretical cell number. HCT116 were plated (96-well) in triplicate at 250 cells per well. Cells were incubated in the presence of vehicle, 1 μ M paclitaxel, or compound. On day 7, CellTiter-Blue (Promega) response was measured and values were normalized to vehicle (100% growth) and paclitaxel (0% growth). For growth assays with inhibitors, $n = 3$ for each concentration with two independent experiments, averaged for Extended Data Table 1, and one experiment shown for graphs of percentage growth versus concentration and time, Fig. 2e and Extended Data Fig. 4a.

Flow cytometric analysis. Cells were plated (6-well) in triplicate at 150,000 cells per ml for 1-day, 2-day and 3-day time points. For the 6-day time point, cells were plated at 35,000 cells per ml and diluted to 150,000 cells per ml with media and compound on day 4. For cell cycle, cells were washed twice with PBS, fixed with 70% ethanol at 4 °C overnight, washed with PBS, and stained with 50 μ g ml⁻¹ propidium iodide (eBioscience) for 1 h at 37 °C. For apoptosis, cells were stained using annexin V-FITC (BD Pharmingen) and 7-AAD (Miltenyi Biotec). Samples were acquired on a BD LSR II and analysed using FlowJo v7.6.5. For the SET-2 differentiation assay, cells were cultured in triplicate with 50 nM CA, 50 ng ml⁻¹ PMA (positive control), or vehicle for 3 days. Cell pellets were collected at 4 °C, washed three times with cold PBS, and stained with anti-CD61-PE (ab91128) or anti-CD41-PerCP (ab134373). For each experiment, $n = 3$ biological replicates with two independent experiments and one shown.

Plasmids, mutagenesis, packaging, transduction, selection and siRNA. 5'-Flag-tagged CDK8 and CDK19 were cloned from pBabe.puro.CDK8.flag⁴⁴ (Addgene 19758) and F-CDK8L (Addgene 24762) into pLVX-EF1alpha-IRES-mCherry and pLVX-EF1alpha-IRES-ZsGreen (Clontech) and transformed into *E. coli* (One Shot Stbl3, Invitrogen). Point mutations were introduced by whole-plasmid PCR (QuikChange II XL Site-Directed Mutagenesis Kit, Agilent). pLVX lentiviral vectors were co-transfected with psPAX and pMD2.G (Addgene) in 293T cells. After 48 h, viral supernatants were collected and passed through a 0.45- μ m filter (Millipore). For transductions, 24-well plates were coated with 500 μ l of 20 μ g ml⁻¹ RetroNectin (Clontech) at 4 °C overnight, blocked with 2% BSA for 30 min, washed with PBS, and 300–500 μ l of viral supernatant was added. The plates were centrifuged (2,000g, 1.5 h) and then set in an incubator. After 2 h, viral supernatant was removed and 500 μ l per well of 200,000 cells per ml was added. After 1–3 days, the cells were expanded and isolated by FACS. Flag-CEBPA (gift from J. Marto), Flag-IRF1 (PlasmID, HMS, HsCD00045286), Flag-IRF8 (PlasmID HMS, HsCD00438293), ETV6-Myc-Flag (Origene, SC118922), CDKN1b-Myc-Flag (Origene, SC117607), and FOSL2-Myc-Flag (Origene, SC110898) were cloned into the Tet-On inducible system pLVX-TRE3G-mCherry or pLVX-TRE3G-ZsGreen (Clontech), transformed into *E. coli* (Stellar Competent Cells, Clontech), packaged into lentiviral vectors and cotransduced with regulator vector pLVX-EF1a-Tet3G. After 1 week of selection with puromycin (1 μ g ml⁻¹) and G418 (400 μ g ml⁻¹), cells were plated in the presence of 100 ng ml⁻¹ doxycycline to assess 7-day growth via Cell-Titer Glo. siRNA against *CEBPA* (Ambion s2888), *IRF1* (Ambion s7501), *ETV6* (Ambion s4867 and s4866), *FOSL2* (Ambion s5345), and *IRF8* (Ambion s7098) or scrambled control (Ambion 4390843) was introduced into cells by electroporation (Amaxa Nucleofector II, Program T-019). After 24 h, cells were plated to assess 3- or 4-day growth via Cell-Titer Glo. Knockdown efficiency was assessed after 24 h by immunoblot or after 48 h by droplet digital PCR (ddPCR). Results shown in Fig. 3g represent a single transduction or a single electroporation. siRNA electroporation and inducible expression cell growth assays were performed 2–6 times. For *ETV6*, two siRNAs were tested with data for siRNA s4867 shown in figures.

Gene expression, gene ontology and GSEA. Leukaemia cells were plated (12-well) in triplicate at 500,000–800,000 cells per ml and incubated in the presence of vehicle or CA (25 nM 3 h for K562, MOLM-14 and MV4;11; 10 nM 24 h for MOLM-14; 25 nM 4 h for SET-2, $n = 3$ for each cell line). Cells were then washed twice with cold PBS, and snap frozen. RNA was isolated (RNeasy Plus Microkit, Qiagen or TRIzol, Life Technologies), processed, and, for K562, MOLM-14 and MV4;11, hybridized to the Human U133 Plus 2.0 microarray (Affymetrix). Microarrays were processed with Bioconductor packages affyQCReport³⁵ for quality control and affy for background correction, summarization, and normalization using rma³⁶. Probe sets present in at least 1 sample (based on affy mas5call) and for which the interquartile range was $>\log_2(1.2)$ were retained for further analysis. The limma Bioconductor package³⁷ was used for differential expression analysis of CA-treated versus DMSO control samples (Benjamini-Hochberg³⁸ adjusted $P < 0.05$). SET-2 and HCT116 gene expression was measured by RNA-seq. SET-2 RNA-seq libraries were prepared and processed using the Ion Torrent workflow. Reads were aligned in two passes, first with rnaStar³⁹ (v.2.3.0e) then with BWA⁴⁰ (v.0.7.5a) for remaining unmapped reads, both using default

parameters. Mapped reads were merged and counted using HTSeq⁴¹ (v.0.5.3p3) with -s yes -m intersection-strict. The Bioconductor package DESeq⁴² was used for DE analysis (FDR < 0.05 and twofold change) and normalization. HCT116 cells were grown to approximately 80% confluence and were treated with either 100 nM CA or DMSO for 3 h ($n = 3$). Cells were then washed twice with cold PBS and scraped into TRIzol reagent (Life Technologies). After collecting the RNA, it was further purified using an RNeasy mini kit (Qiagen) with an on-column DNase I digestion. Libraries for Illumina sequencing were generated via the Illumina TruSeq stranded mRNA prep kit. Samples were run in a single lane on an Illumina HiSeq 2000 sequencer with a single read flow cell using 1 \times 50-bp reads and a 6-cycle index read. Reads were mapped to the hg19 reference genome using Tophat2 v.2.0.6 with custom settings including the setting of -library-type fr-firststrand to appropriately account for the stranded nature of the protocol. HTSeq v.0.6.1 was used to obtain read counts over annotated genes and differentially expressed genes were called by DESeq v.1.10.1 with a padj value of less than 0.01. Counts were normalized for GSEA using the limma voom function⁴³. Expression data for the I-BET151 comparison were downloaded from ArrayExpress (https://www.ebi.ac.uk/arrayexpress, accession E-MTAB-774) and processed files used as is. Gene lists were submitted to the DAVID web server (http://david.abcc.ncifcrf.gov) for functional annotation⁴⁴. GSEA version 2.09 (ref. 19) was carried out using signal-to-noise on natural values as the metric. Signatures included curated gene sets (C2, v.3) downloaded from the Broad's MSigDB as well as signatures curated from in-house and published data sets.

ChIP-seq. Untreated cells or cells treated with CA (25 nM, 6 h), iBET-151 (500 nM, 6 h) or vehicle were crosslinked for 10 min at room temperature by addition of one-tenth of the volume of formaldehyde solution (11% formaldehyde, 50 mM HEPES, pH 7.4, 100 mM NaCl, 1 mM EDTA and 0.5 mM EGTA) to the media followed by 5 min quenching with 125 mM glycine. For CDK8 and MED1 chromatin immunoprecipitations, cells were instead centrifuged, resuspended in serum-free media, and crosslinked at room temperature by addition of an equal volume of 2% formaldehyde in serum-free media for 10 min followed by quenching with 125 mM glycine for 5 min. Cells were then washed twice with cold PBS and snap frozen. ChIP was performed essentially as previously described². In brief, cells were lysed with lysis buffer 1 (50 mM HEPES, pH 7.4, 140 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% NP-40 and 25% Triton X-100) and washed with lysis buffer 2 (10 mM Tris-HCl, pH 8.0, 200 mM NaCl, 1 mM EDTA and 0.5 mM EGTA). For H3K4me3, H3K27me3, H3K27ac, H3K4me1 and pol II, the nuclei were resuspended in 10 mM Tris-HCl, pH 8.0, 100 mM NaCl, 1 mM EDTA, pH 8.0, 0.5 mM EGTA, 0.1% Na-deoxycholate and 0.2% SDS, sheared for 2 min (Branson S220D sonifier, pulse, 0.7 s on, 1.3 s off, 12–14 W) on wet ice, and then Triton X-100 was added to 1% (v/v). For MED1 and CDK8, the nuclei were resuspended in 50 mM Tris-HCl, pH 7.5, 140 mM NaCl, 1 mM EDTA, 1 mM EGTA, 0.1% SDS and 1% Triton X-100 then sheared for 4 min (pulse, 0.7 s on, 1.3 s off, 10–12 W) on wet ice. Sonicated lysates were cleared and incubated overnight at 4 °C with Protein G magnetic Dynabeads (50 μ l) pre-bound with the indicated antibodies (5 μ g). Beads were washed with sonication buffer, sonication buffer with 500 mM NaCl, LiCl wash buffer (20 mM Tris-HCl, pH 8.0, 1 mM EDTA, 250 mM LiCl, 0.5% NP-40, 0.5% sodium deoxycholate) and TE. Bound complexes were eluted with 50 mM Tris-HCl, pH 8.0, 10 mM EDTA, 1% SDS at 65 °C and reverse crosslinked at 65 °C. RNA and protein were digested using RNase A and proteinase K, respectively, and DNA was purified using Qiagen MinElute columns. Libraries for Illumina sequencing were prepared using the Illumina TruSeq ChIP Sample Preparation kit with the following exceptions. After end-repair and A-tailing, ChIP DNA or whole-cell extract DNA was ligated to Illumina RNA adaptors with unique indices. Alternatively, libraries were prepared using the KAPA Hyper Prep Kit for Illumina and ligated to unique Bio Scientific NEXTflex barcode adaptors. Following ligation, libraries were amplified with 16–18 cycles of PCR and were then size-selected using a 2% gel cassette in the Pippin Prep System from Sage Science. For histone modifications and RNA pol II, DNA fragments of size 200–500 bp were captured. For CDK8 and MED1, DNA fragments of size 200–450 bp were captured. Libraries were quantified by qPCR using the KAPA Biosystems Illumina Library Quantification kit. Libraries with distinct indexes were then combined in equimolar ratios and run together in a lane on the Illumina HiSeq 2500 for 40 bases in single read mode.

ChIP-seq data analysis. ChIP-seq data sets were aligned using Bowtie (v.0.12.8)⁴⁵ to build version NCBI37/HG19 of the human genome (-n 1 -m 1-best-strata). Duplicate reads were removed using Picard tools (v.1.88). For CDK8, peaks were called with both SPP⁴⁶ and MACS v.1.4 (ref. 47) using default significance cut-off values. SPP cross-correlation analysis was used for both quality control⁴⁸ and to set the strand shift parameter for MACS. Regions of interest identified by both peak callers were retained and merged. Regions overlapping $>70\%$ with RepeatMasker regions (downloaded 16 November 2012 from UCSC) were excluded from further analysis. Retained regions were annotated by overlap with RefSeq genes (genomic

coordinates downloaded from UCSC refgene table Apr. 26, 2013) using bedtools⁴⁹. Retained regions were assigned to one of the following categories: (1) promoter = transcription start site (TSS) – 500 bp to TSS + 200 bp, (2) body = TSS + 201 bp to TES, (3) proximal enhancer = TSS – 5 kb to TSS – 501 bp, and (4) 3' untranslated region (UTR) = TES + 1 bp to TES + 5 kb. All other regions were termed 'desert' hits. Any gene satisfying the overlap criteria was included in the corresponding category. Travelling ratios were calculated essentially as described⁵⁰. In brief, mapped read coordinates were first extended 3' to 200 bases to capture the full fragment coverage. The RefSeq coordinates used for annotation were then used to count extended pol II reads falling in the range of TSS – 30 bp to TSS + 300 bp and those falling in the remainder of the gene body (TSS + 301 to TES). Very short transcripts (<630 bp) were excluded, as were cases with very low counts in both regions. Input reads were subtracted and counts were scaled to reads per kilobase. Transcripts sharing identical TSS and TES coordinates were represented a single time in the count statistics. ChIP-seq tracks were smoothed by calculating the density per million mapped reads in 300 bp bins at 50 bp intervals and were visualized using Integrative Genomics Viewer. ChIP-seq density maps were generated using ngsplot⁵¹ (v.2.08). Heatmap of semi-supervised clustering in Fig. 1a of total signal on CDK8 positive regions was carried out as follows: (1) peaks were individually identified for each of the 6 ChIPs using MACS2 at default *P* value cutoff; (2) all peaks were combined and merged into non-redundant regions using mergeBed (-d 0); (3) within each unique region, ChIP reads were counted and matched input reads were subtracted after scaling each to million mapped reads; (4) clusters were grouped by ChIPs represented in a given region into 64 categories in the following order: H3K4me1, H3K27ac, pol II, MED1 and BRD4; (5) each group was ordered by decreasing CDK8 signal per region; and (6) ChIP samples were clustered by Euclidean distance of ChIP signal per region after median centring and normalization. A similar approach was used for BRD4 and CDK8 ChIPs in MOLM-14 cells treated with DMSO or I-BET151. In this case, non-promoter-associated regions in which I-BET151 treatment reduced BRD4 signal >2-fold were ordered by log₂ fold-change.

Irreproducibility discovery rate analysis. Reproducibility of two independent H3K27ac ChIP-seq experiments carried out in cells treated with either DMSO or CA for 3 h was assessed according to the pipeline developed for the ENCODE project (<https://sites.google.com/site/anshulkundaje/projects/idr>)⁵². Irreproducibility discovery rate (IDR) was determined as recommended on peaks called by SPP⁴⁶ at FDR < 0.5. At this threshold, SPP reported between 180,000 and 300,000 peaks, depending on the exact combination of sample and input, most of which are expected to be noise. Under both treatment conditions, the number of high-confidence peaks (IDR threshold < 0.01 for true replicates and pseudo-replicate self-consistency tests and < 0.0025 for pseudo-replicate pooled-consistency analysis) identified based on signal value in the replicates and pseudo-replicates was within the recommended twofold range, indicating good reproducibility. The number of peaks with IDR < 0.01 in the true replicates was used to make the final selection of distinct, non-chrM pooled replicate peaks. Regions within 200 nucleotides of each other were merged to generate the final peaks list. The same approach was used to determine reproducible peaks in two independent BRD4 and CDK8 ChIP experiments in MOLM-14 cells treated with DMSO or I-BET151.

Identification of SEs. MED1 signal was measured in active enhancers (that is, regions enriched in both H3K4me1 and H3K27ac) after extending MED1 ChIP-seq reads 100 bases in a strand-aware fashion. Enhancer regions were sorted based on their MED1 signal and the inflection point of the curve determined. Enhancers with MED1 signal above the inflection point were retained as SEs⁵. In a separate approach, using only the MED1 ChIP-seq data and the ROSE software from the Young laboratory¹, we found >80% agreement with our previous assignment of MED1 SEs. ROSE was used thereafter to identify SEs using BRD4, H3K27ac (±CA, 3 h), and CDK8 ChIP-seq on peaks called by MACS 1.4. For K562 and HCT116, H3K27ac ChIP samples and their matched inputs were downloaded from the ENCODE project repository at UCSC (sample identifiers and references in Supplementary Table 1). For HCT116, CDK8 ChIP-seq data and matched input was downloaded from GSE38258 (ref. 53). SE-associated genes were assigned to the nearest expressed transcript, based on H3K27ac signal in a 500-nucleotide window centred on the TSS¹. For Extended Data Fig. 1e, we normalized each experiment's signal (after adjusting to million mapped reads and subtracting input signal) to show values from independent ChIP-seq experiments on a common scale. Normalized signal for each enhancer, *x*, is thus $(x - \text{minimum}) / (\text{maximum} - \text{minimum})$. Each ChIP-seq experiment yielded different numbers of enhancer regions so we mapped each experiment's enhancer ranks to [0,1] by calculating $(\text{rank} - 1) / (\text{maximum rank} - 1)$.

RNA levels, ddPCR and qRT-PCR. Total RNA was isolated from 500,000 MOLM-14 cells (RNeasy Plus Mini Kit, Qiagen) and quantified by Nanodrop. mRNA was subsequently isolated (Dynabeads mRNA Purification Kit, Life Technologies) and quantified by Nanodrop. For ddPCR, total RNA was

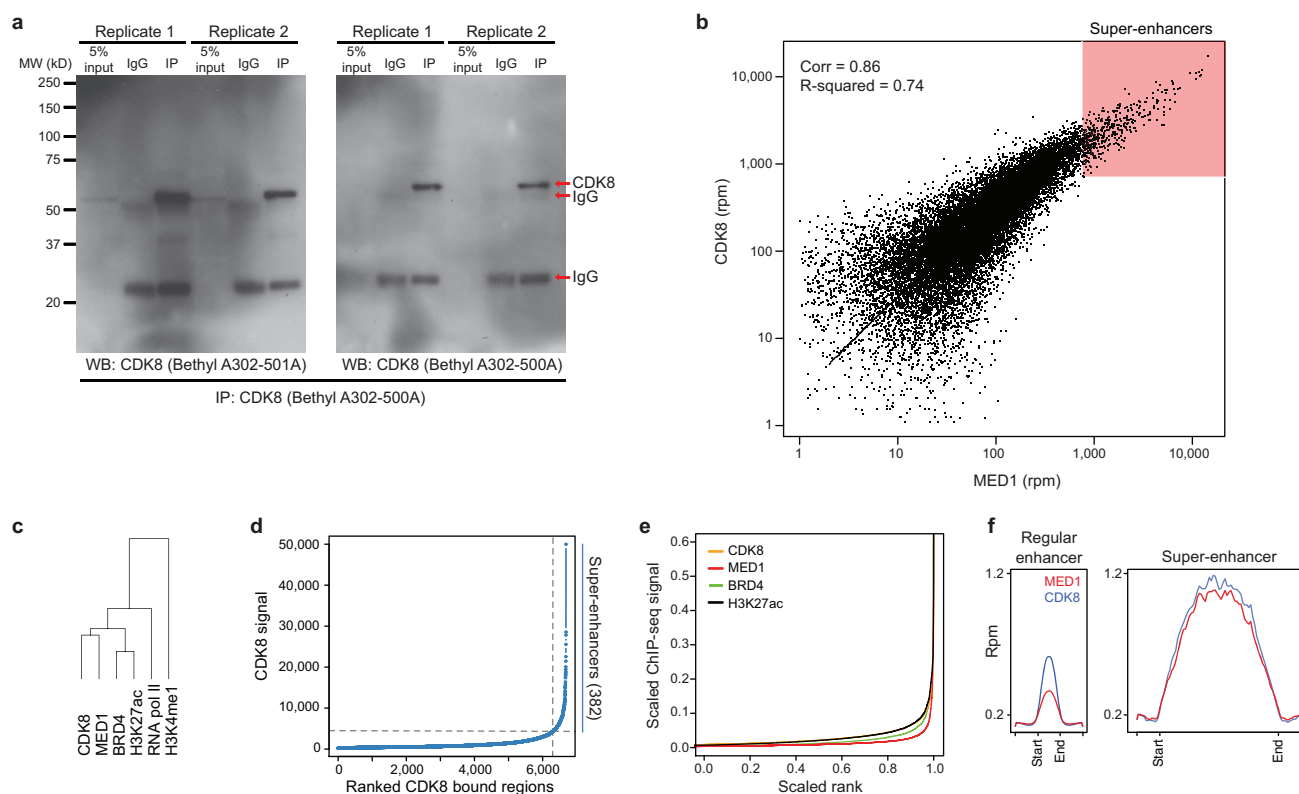
reverse-transcribed into cDNA (High Capacity cDNA Reverse Transcriptase Kit, Applied Biosystems) and used (ddPCR Supermix for Probes, no dUTP, Bio-Rad 186-3024) with TaqMan FAM probes for genes of interest and *ACTB* (VIC) as the reference gene. Droplets were generated in the QX200 Droplet Generator, thermocycled, and read on the QX200 Droplet Reader. Total RNA per cell was measured by isolating total RNA from 10⁶ cells using the mirVana miRNA Isolation Kit (Life Technologies) and quantifying by Nanodrop. The difference in copy numbers of specific mRNAs before and after treatment (Fig. 3e) was determined relative to copies of *ACTB* mRNA per cell. Probes used (Life Technologies): CEBPA (Hs00269972_s1), ETV6 (Hs00231101_m1), IRF1 (Hs00971960_m1), IRF8 (Hs00175238_m1), RREB1 (Hs01002873_m1), CDKN1B (Hs01597588_m1), GFII1 (Hs00382207_m1), JARID2 (Hs01004460_m1), BHLHE40 (Hs01041212_m1), and *ACTB* (4325788). qRT-PCR for checking siRNA knockdown was performed with iTaq Universal Probes Supermix (Bio-Rad), *n* = 3, or by ddPCR.

In vivo studies. Studies were performed at Charles River Laboratories (CRL) and Dana Farber Cancer Institute (DFCI) where indicated and approved by Harvard University and each institution's respective animal care and use committee. For pharmacokinetic studies, serial blood samples from 7-week-old male CD-1 mice (*n* = 3 per time point) were collected (no blinding) into K₂EDTA tubes, centrifuged, transferred into 96-well plates (matrix tubes), stored at –20 °C, and analysed by liquid chromatography–tandem mass spectrometry (LC–MS/MS) (*in vivo* studies performed at CRL). Study size was determined by the need for three blood samples per time point with three blood samples collected per mouse. The MV4;11 xenograft model were performed as previously described²¹ (*in vivo* studies performed at DFCI) Two-million MV4;11-mCLP cells were injected into the tail vein of 7-week-old female non-obese diabetic–severe combined immunodeficient (NOD–SCID) *Il2rg*^{–/–} (NSG) mice (The Jackson Laboratory) and tumour burden was assessed by bioluminescence imaging (BLI) using an IVIS Spectrum system (Caliper Life Sciences). Seven days after injection, leukaemia establishment was documented by BLI and mice were assigned to groups to achieve a similar mean BLI and treated intraperitoneally with vehicle (20% hydroxypropyl-β-cyclodextrin) or CA once daily for 15 days. After 30 days, blood counts were obtained (Hemavet 950 F, Drew Scientific) and spleen, femur and peripheral blood cells were collected and analysed by flow cytometry (LSR Fortessa, BD Biosciences) from three mice per group. The mice and a portion of the spleen were preserved in bouins after body cavities were opened and visceral organs exposed. Samples from all organs were then dissected and placed in nine cassettes per mouse. Tissues were paraffin embedded, sectioned at 6 μm and stained with haematoxylin and eosin. Survival was measured as the time from therapy initiation until moribund state. We selected 11 mice per group to match previous survival analysis in the model²¹ (*n* = 8) and to have 3 additional mice per group for disease burden comparison. Blinding was only done for histopathology analysis. For the SET-2 xenograft model (*in vivo* studies performed at CRL), 8–12-week-old female SCID Beige mice (Charles River) were injected subcutaneously in the flank with 10⁷ SET-2 cells in 50% matrigel (0.2 ml per mouse). When tumours reached an average size of 80–120 mm³, mice were assigned to groups to achieve a similar mean tumour size and treatment commenced without blinding. Tumour volumes were measured using calipers and calculated as $(\text{width}^2 \times \text{length}) / 2$. Percentage tumour growth inhibition was calculated as $(\text{vehicle} - \text{treatment}) / (\text{vehicle} - \text{initial}) \times 100$. We selected 10 mice per group to safeguard against the IACUC requirement to stop dosing a group if >10% mortality occurs. For safety testing (*in vivo* studies performed at DFCI), 8-week-old female CD-1 mice were treated once daily without blinding for 15 days and weighed daily. Two hours after the last dose, blood counts were obtained and blood chemistry was analysed. Three mice per group were selected as a minimum for comparison. For STAT1-pS727 inhibition, 6–10-week-old female C57BL/6 mice were treated once daily for 2 days (*in vivo* studies performed at CRL, not blinded). One hour after the second dose, natural killer cells were isolated by dissociation of splenocytes from isolated spleens, lysis of erythrocytes, and isolation of DX5⁺ cells (MiniMACS CD49b, Miltenyi Biotec) and analysed by immunoblot and densitometry (ImageJ, STAT1-pS727 level normalized to β-actin). We selected three mice per group as a minimum for comparison. Statistical analyses were performed using GraphPad Prism 6.0. For *P* value determinations, two-way or one-way ANOVA was used with Dunnett's multiple comparison testing and *P*-value adjustment. Dotted purple lines were from the Mouse Phenome Database 22903 (The Jackson Laboratory). No statistical methods were used to predetermine sample size, and experiments were not randomized.

23. Tahirov, T. H. et al. Crystal structure of HIV-1 Tat complexed with human P-TEFb. *Nature* **465**, 747–751 (2010).

24. Knuesel, M. T., Meyer, K. D., Donner, A. J., Espinosa, J. M. & Taatjes, D. J. The human CDK8 subcomplex is a histone kinase that requires Med12 for activity and can function independently of mediator. *Mol. Cell. Biol.* **29**, 650–661 (2009).

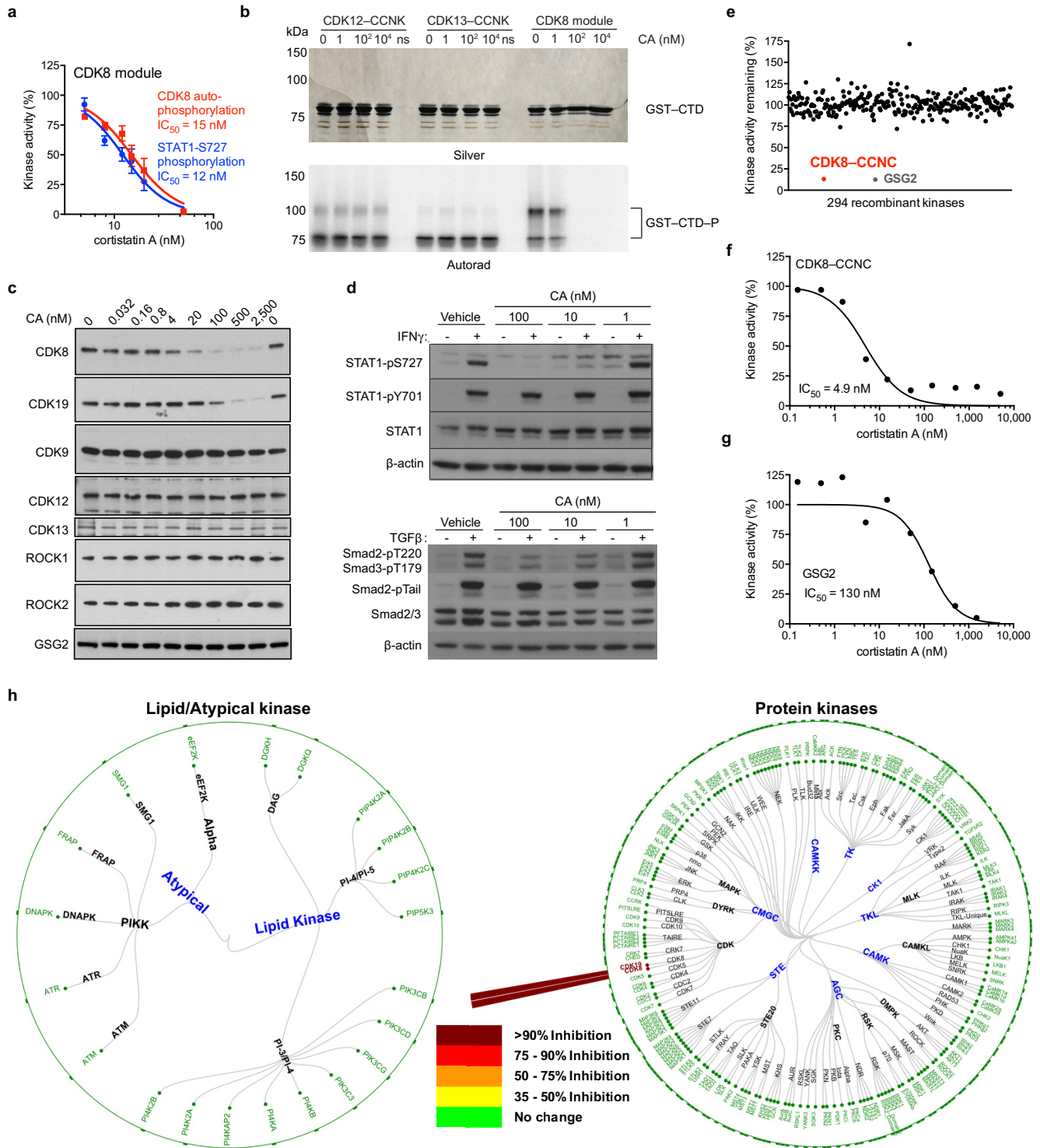
25. Okerberg, E. *et al.* Profiling native kinases by immuno-assisted activity-based profiling. *Curr Protoc Chem Biol* **5**, 213–226 (2013).
26. Hutterer, C. *et al.* A novel CDK7 inhibitor of the pyrazolotriazine class exerts broad-spectrum antiviral activity at nanomolar concentrations. *Antimicrob. Agents Chemother.* **59**, 2062–2071 (2015).
27. Schneider, E. V. *et al.* The structure of CDK8/CycC implicates specificity in the CDK/Cyclin family and reveals interaction with a deep pocket binder. *J. Mol. Biol.* **412**, 251–266 (2011).
28. Kabsch, W. Integration, scaling, space-group assignment and post-refinement. *Acta Crystallogr. D* **66**, 133–144 (2010).
29. Vagin, A. & Teplyakov, A. Molecular replacement with MOLREP. *Acta Crystallogr. D* **66**, 22–25 (2010).
30. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
31. Dodson, E. J., Winn, M. & Ralph, A. Collaborative computational project, number 4. Providing programs for protein crystallography. *Methods Enzymol.* **277**, 620–633 (1997).
32. Vagin, A. A. *et al.* REFMAC5 dictionary: organization of prior chemical knowledge and guidelines for its use. *Acta Crystallogr. D* **60**, 2184–2195 (2004).
33. The PyMOL molecular graphics system v. 1.3r1 (Schrödinger, LLC, 2010).
34. Firestein, R. *et al.* CDK8 is a colorectal cancer oncogene that regulates β -catenin activity. *Nature* **455**, 547–551 (2008).
35. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).
36. Rafael, R. A. *et al.* Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* **31**, e15 (2003).
37. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
38. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
39. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
40. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
41. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
42. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
43. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
44. da Huang, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4**, 44–57 (2009).
45. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. *Genome Biol.* **10**, R25 (2009).
46. Kharchenko, P. V., Tolstorukov, M. Y. & Park, P. J. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nature Biotechnol.* **26**, 1351–1359 (2008).
47. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
48. Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22**, 1813–1831 (2012).
49. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
50. Rahl, P. B. *et al.* c-Myc regulates transcriptional pause release. *Cell* **141**, 432–445 (2010).
51. Shen, L., Shao, N., Liu, X. & Nestler, E. ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. *BMC Genomics* **15**, 284 (2014).
52. Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.* **5**, 1752–1779 (2011).
53. Galbraith, M. D. *et al.* HIF1A employs Cdk8-mediator to stimulate RNAPII elongation in response to hypoxia. *Cell* **153**, 1327–1339 (2013).



Extended Data Figure 1 | CDK8 ChIP-seq defines SE-associated genes.

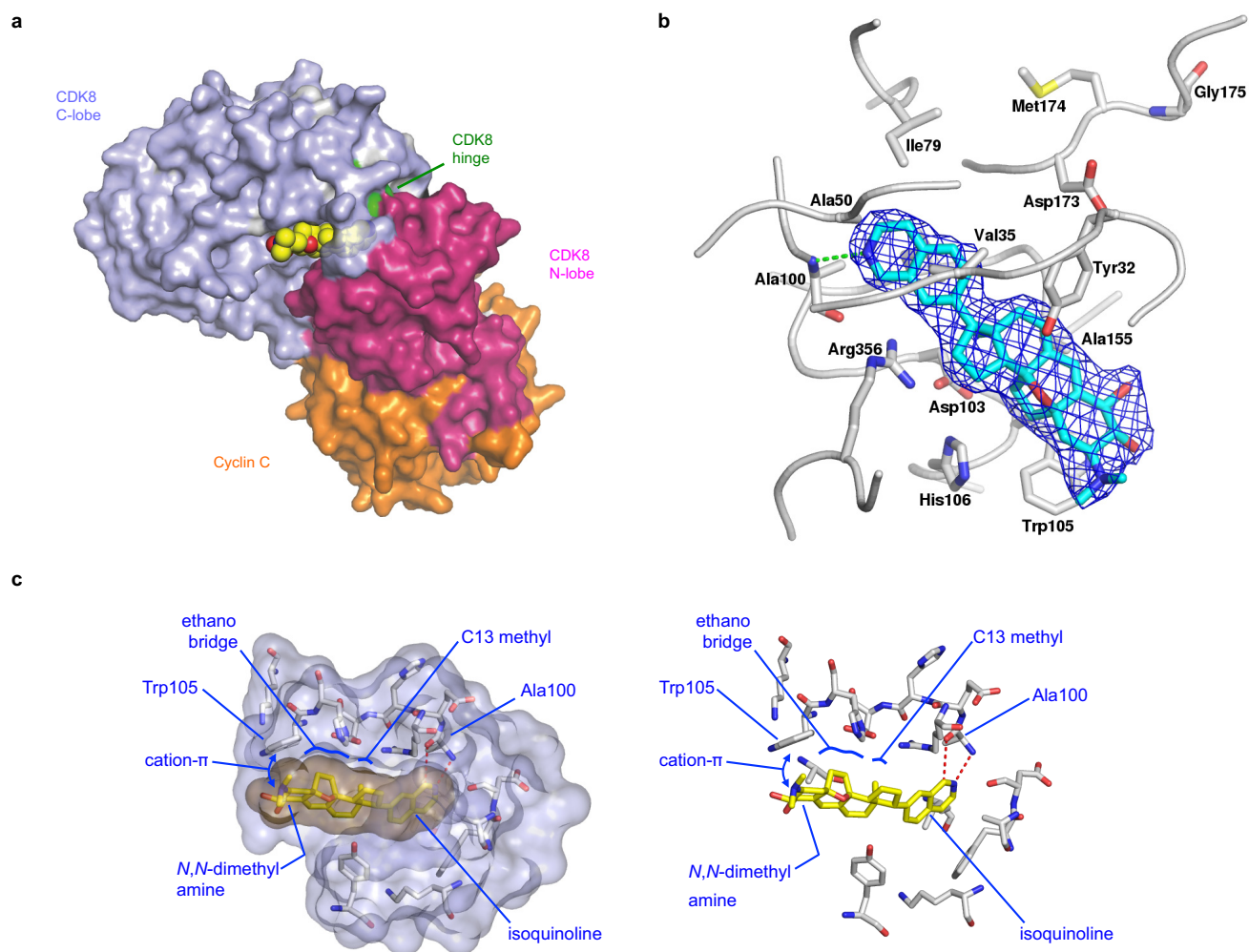
a, The antibody used for CDK8 ChIP-seq (Bethyl A302-500A) was validated by immunoprecipitation (IP) and western blot (WB). Immunoprecipitation was conducted with Bethyl A302-500A (2 μ g) on MOLM-14 whole-cell extract, and western blot was performed on split immunoprecipitation lysate or 5% input with either anti-CDK8 Bethyl A302-501A (left), anti-CDK8 Bethyl A302-500A (right), or normal rabbit IgG (CST, 2729), experiment performed once. **b**, MED1 and CDK8 density is highly correlated on active enhancer regions marked by H3K4me1 and H3K27ac (correlation = 0.86, R^2 = 0.74) in

MOLM-14 cells. The pink box represents SEs. **c**, Hierarchical clustering dendrogram of CDK8, MED1, BRD4, H3K27ac, RNA pol II and H3K4me1 ChIP-seq signal. **d**, Distribution of CDK8 signal with input subtracted across CDK8 bound regions. Regions to the right of inflection point are considered SEs. **e**, Distribution of CDK8, MED1, BRD4 and H3K27ac signal across putative enhancer regions. Regions to the right of the distribution inflection point are considered SEs. **f**, ChIP-seq profile plots centred around MED1-defined SE and regular enhancer regions. Flanking regions are 2.5 kb.



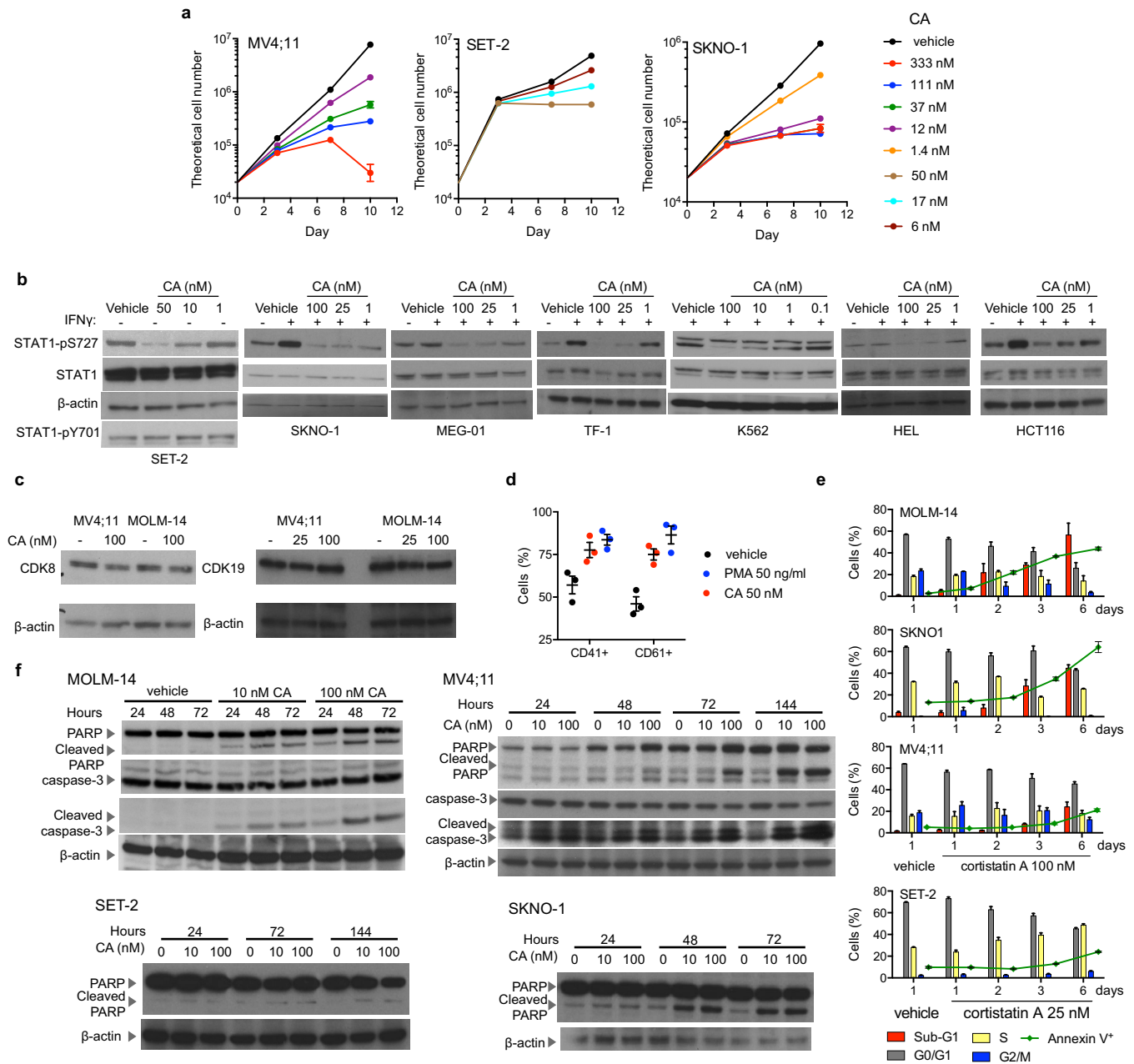
Extended Data Figure 2 | CA inhibition of and binding to CDK8. **a**, CA inhibition of CDK8 module phosphorylation of CDK8 and STAT1-S727 substrate (mean \pm s.e.m., $n = 3$ biological replicates, one of two experiments shown, autorad in Supplementary Fig. 1). **b**, CA inhibition *in vitro* of CDK8 module activity but not CDK12-CCNK or CDK13-CCNK activity up to 10 μ M. Equal amounts (silver stain) of GST-CTD were used as the substrate in *in vitro* kinase assays. The amount of each kinase used was empirically determined to give approximately the same GST-CTD signal under the assay conditions. GST-CTD-P, phosphorylated GST-CTD; ns, no substrate (kinase only). One of four experiments shown. **c**, Immunoblot showing that CA selectively and dose-dependently inhibits capture of native CDK8 ($IC_{50} \approx 10$ nM) and CDK19 ($IC_{50} \approx 100$ nM) from MOLM-14 lysates but does

not inhibit capture of CDK9, CDK12, CDK13, ROCK1, ROCK2 or GSG2. One of two experiments shown, full scan in Supplementary Fig. 1. **d**, Immunoblots showing CA inhibition of CDK8-dependent IFN- γ -stimulated STAT1-S727 phosphorylation in MOLM-14 cells and CA inhibition of TGF- β -stimulated Smad2-T220 and Smad3-T179 phosphorylation in HaCaT cells ($IC_{50} < 100$ nM). One of two experiments shown, full scan in Supplementary Fig. 1. **e**, *In vitro* kinase activity profiling (mean for kinase reaction, $n = 2$ biological replicates, experiment performed once). **f**, **g**, CA dose-dependent inhibition of CDK8-CCNC complex ($IC_{50} = 5$ nM) (**f**) and GSG2 ($IC_{50} = 130$ nM) (**g**) as measured in **e** ($n = 1$, experiment performed once). **h**, Dendrogram representation of results shown in Fig. 2c for 1 μ M CA.



Extended Data Figure 3 | CA-CDK8-CCNC ternary complex. **a**, The 2.4 Å crystal structure of the human CA-CDK8-CCNC ternary complex shown as a Corey-Pauling-Koltun (CPK) model. **b**, CA and neighbouring protein side chains are shown as a stick model coloured according to the chemical atom type (CA in cyan, CDK8-CCNC in grey, N in blue, O in red and S in yellow). CA is shown superimposed with the refined $2F_o - F_c$ electron density map

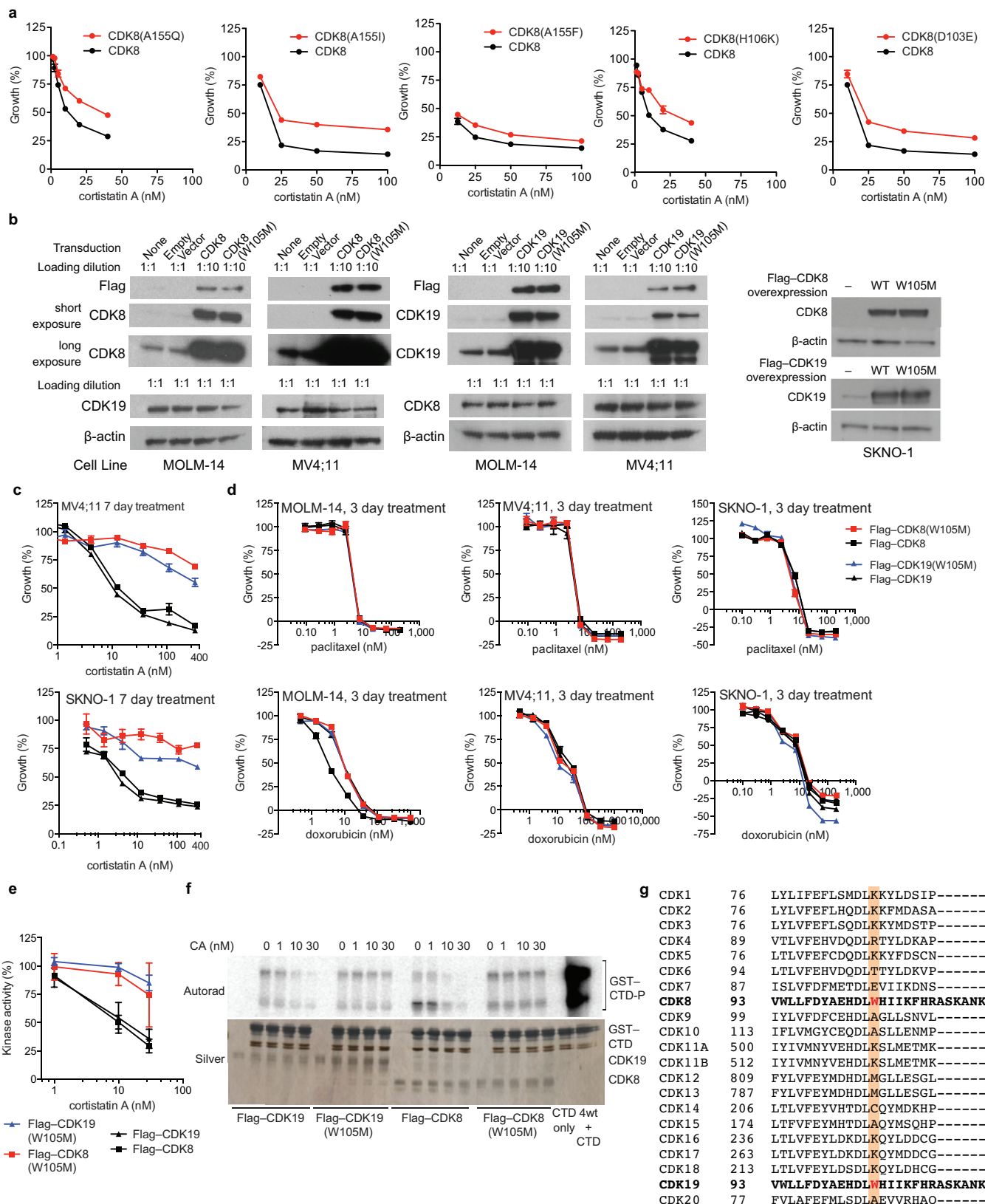
contoured at 1.0σ . Hydrogen bonds are indicated as green dotted lines. **c**, A portion of the CA-CDK8-CCNC crystal structure showing the CA binding pocket of CDK8 (with and without a semi-transparent surface; CA in gold, CDK8 in grey) with certain residues and CA in stick representation. Dotted red lines indicate H-bonds. Key residues and binding elements are labelled.



Extended Data Figure 4 | Antiproliferative activity of CA and I-BET151.

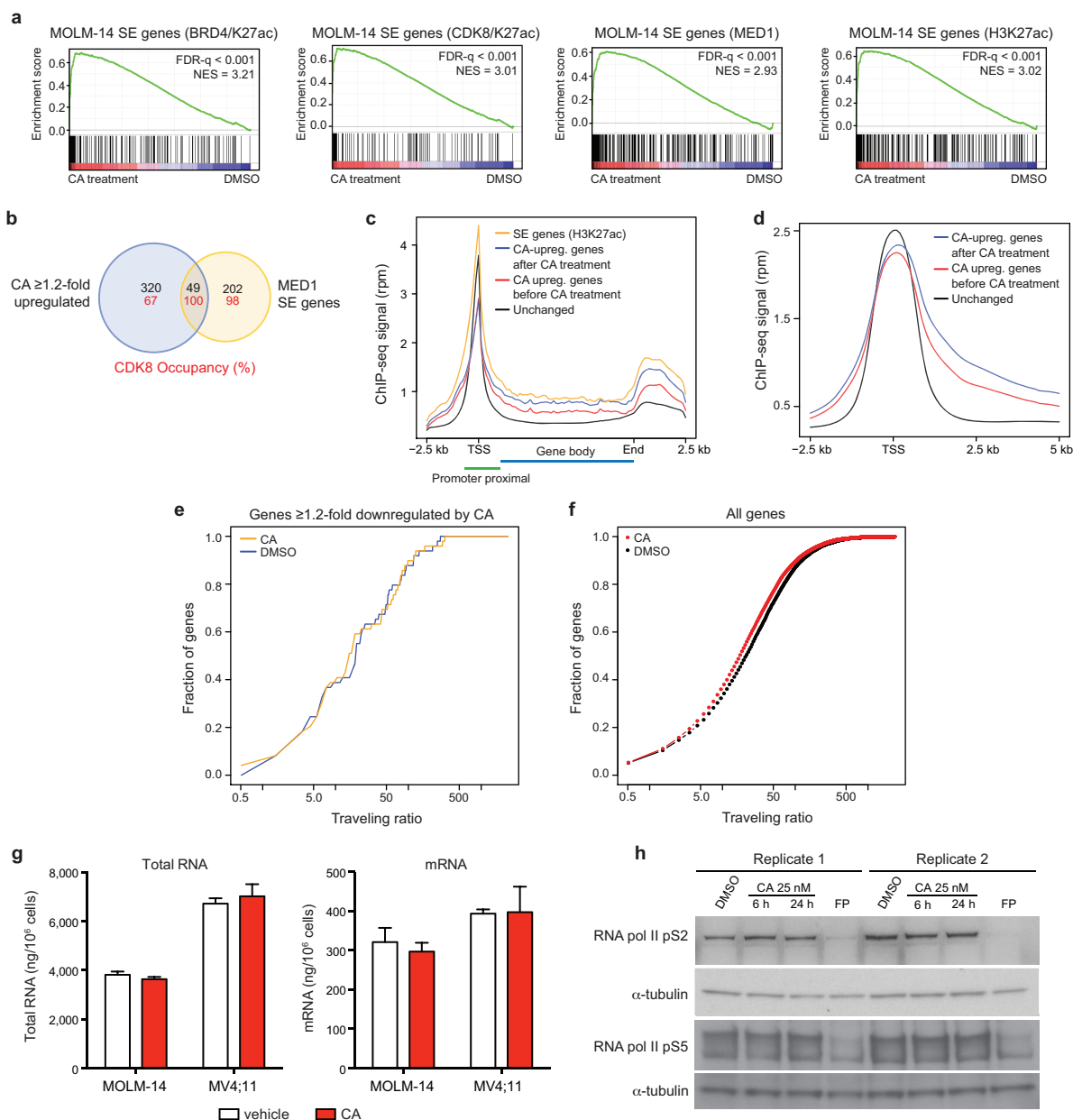
a, Plots showing antiproliferative activity of CA over time for selected sensitive cell lines and concentrations (mean \pm s.e.m., $n = 3$ biological replicates, one of two experiments shown). **b**, Immunoblots showing that CA inhibits CDK8-dependent IFN- γ -stimulated STAT1-pS727 phosphorylation equally well in cells sensitive or insensitive to the antiproliferative activity of CA (one of two experiments shown, full scan in Supplementary Fig. 1). **c**, Immunoblots showing CDK8 and CDK19 levels after 24 h CA treatment in sensitive cell lines MV4;11 and MOLM-14 (one of two experiments shown, full scan in

Supplementary Fig. 1). **d**, CD41 and CD61 (vehicle versus CA, $P = 0.04$ and 0.005 , respectively, two-tailed t -test) on SET-2 cells after 3 days of indicated treatment (mean \pm s.e.m., $n = 3$ biological replicates, one of two experiments shown). Phorbol 12-myristate 13-acetate (PMA) was used as positive control. **e**, DNA content and annexin V staining of indicated cell lines after treatment with CA (mean \pm s.e.m., $n = 3$ biological replicates, one of two experiments shown). **f**, Immunoblots of CA dose- and time-dependent induction of PARP and caspase-3 cleavage for indicated cell lines (one of two experiments shown, full scan in Supplementary Fig. 1).



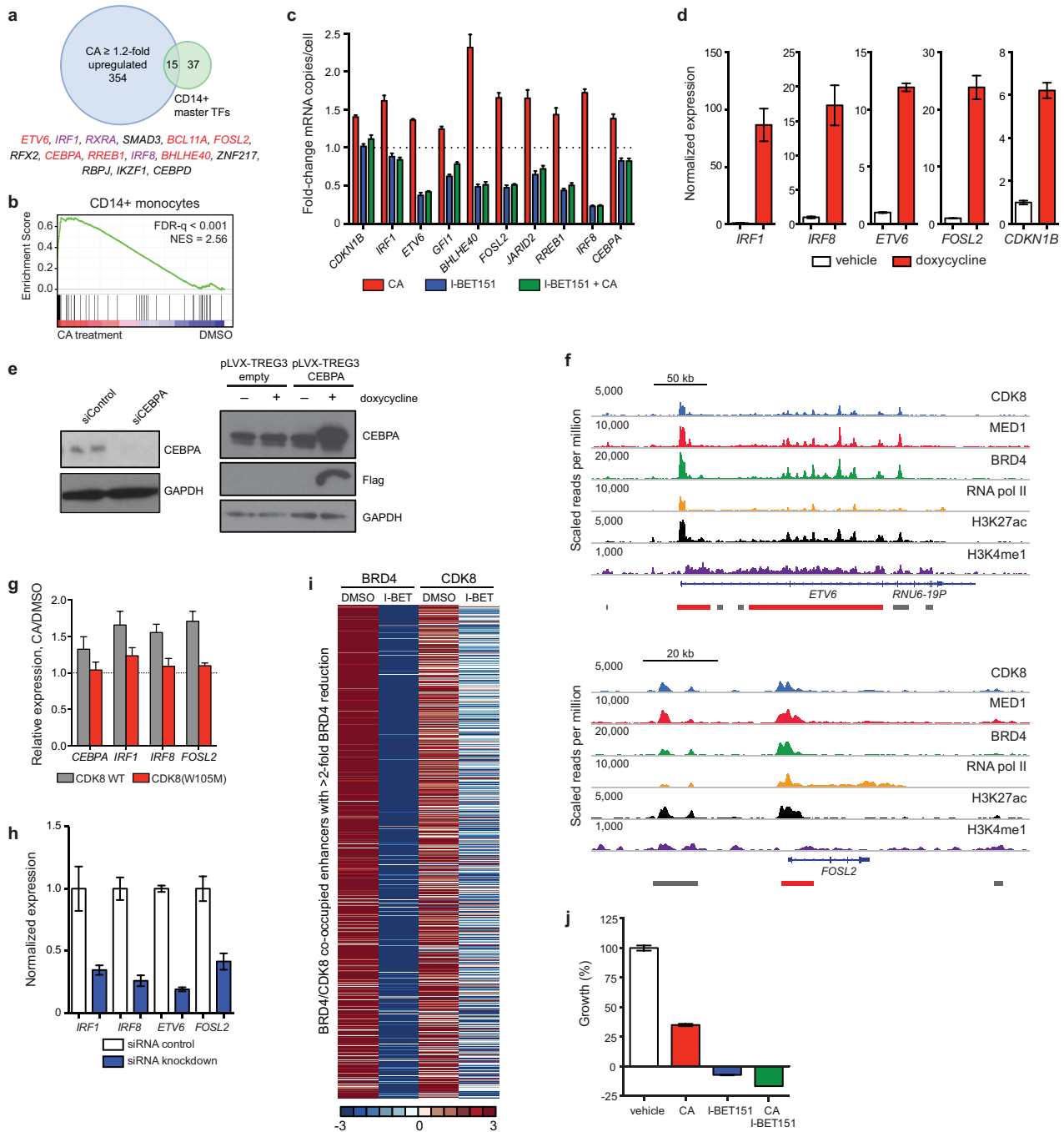
Extended Data Figure 5 | Mediator kinases mediate the antiproliferative activity of CA. **a**, We evaluated point mutations to CDK8 residues lining the CA-binding pocket: Ala155, His106, Asp103 and Trp105. Expression of CDK8 mutants A155I, A155F, A155Q, H106K and D103E in MOLM-14 cells afforded only modest desensitization to CA. Differential sensitivity of MOLM-14 cells to CA after expression of indicated mutant Flag-CDK8 proteins (mean \pm s.e.m., $n = 3$ biological replicates, experiment performed once). **b**, Immunoblots showing that Flag-CDK8 or Flag-CDK19 and Flag-CDK8(W105M) or Flag-CDK19(W105M) are expressed at similar levels in MOLM-14, MV4;11 and SKNO-1 cells (experiment performed once, full scan in Supplementary Fig. 1). **c**, Differential sensitivity of MV4;11 and SKNO-1 cells to CA after expression of Flag-CDK8, Flag-CDK19, Flag-CDK8(W105M) and Flag-CDK19(W105M), legend as in **d** (mean \pm s.e.m., $n = 3$ biological replicates, one of two experiments shown). **d**, Control showing that expression of Flag-CDK8(W105M) or Flag-CDK19(W105M) in MOLM-14, MV4;11 and SKNO-1 cells does not confer resistance to

antiproliferative agents paclitaxel and doxorubicin (mean \pm s.e.m., $n = 3$ biological replicates, one of two experiments shown). **e**, Purified Flag-CDK8(W105M) and Flag-CDK19(W105M) remain catalytically active for phosphorylation of CTD *in vitro* but are resistant to inhibition by CA (mean \pm s.e.m., $n = 3$ biological replicates, experiment performed once). **f**, Representative autorad and silver stain images supporting quantification shown in **e**. **g**, Sequence alignment of human CDKs. Sequence alignment was performed on segments of CDK1-20 using Clustal Omega. The unique Trp105 residue in CDK8 and CDK19 is highlighted in red, and is absent from other CDKs (orange box). UniProt Knowledgebase entries: CDK1, P06493; CDK2, P24941; CDK3, Q00526; CDK4, P11802; CDK5, Q00535; CDK6, Q00534; CDK7, P50613; CDK8, P49336; CDK9, P50750; CDK10, Q15131; CDK11A, Q9UQ88; CDK11B, P21127; CDK12, Q9NYV4; CDK13, Q14004; CDK14, O94921; CDK15, Q96Q40; CDK16, Q00536; CDK17, Q00537; CDK18, Q07002; CDK19, Q9BWU1; CDK20, Q8IZL9.



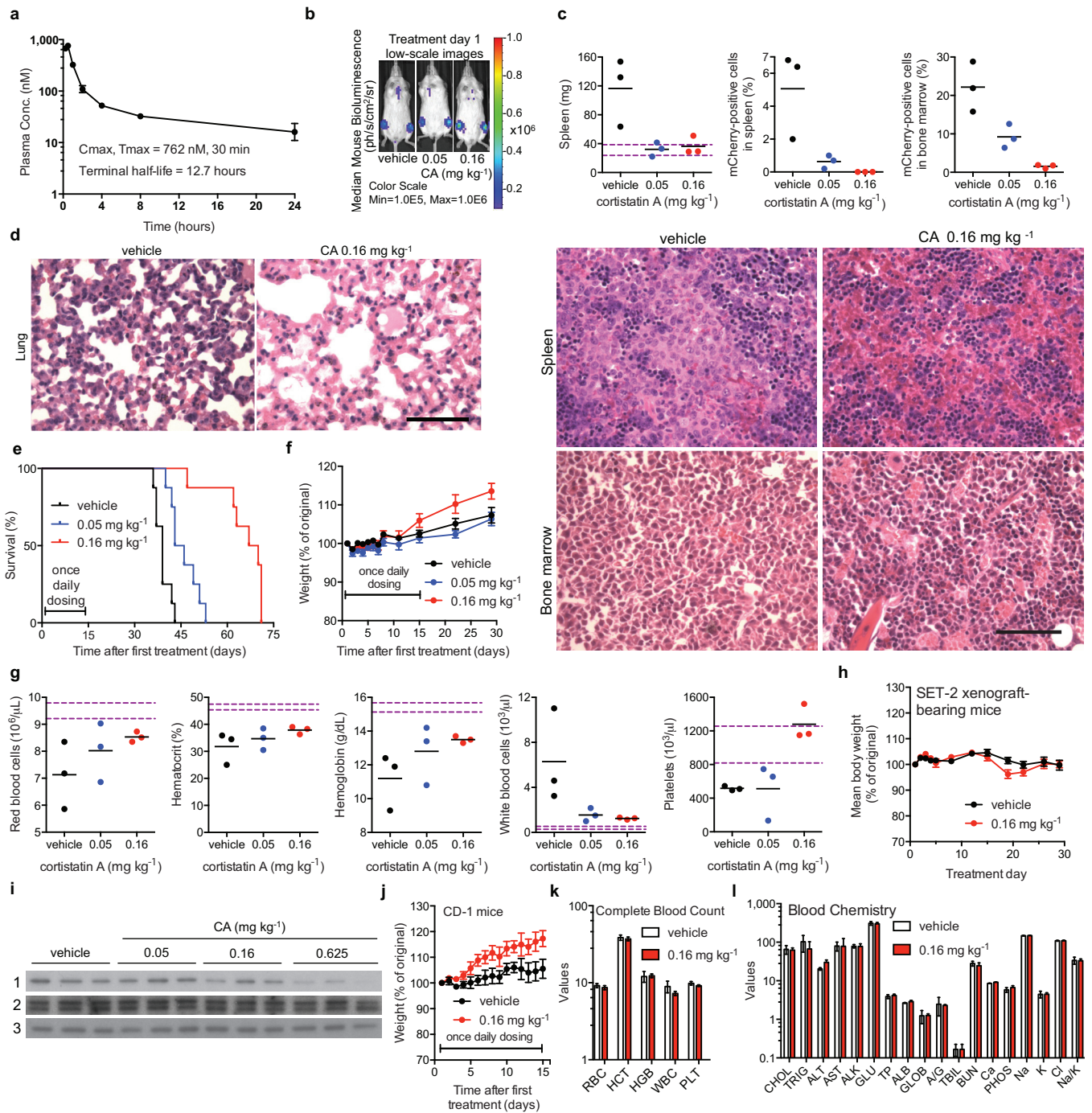
Extended Data Figure 6 | CA disproportionately affects expression of SE genes in MOLM-14 cells. **a**, GSEA plots showing positive enrichment of SE-associated genes, defined by ChIP-seq signal for indicated factors, with 3 h CA treatment in MOLM-14 cells (differential expression versus DMSO controls). **b**, Venn diagram showing the overlap between SE genes and genes upregulated ≥ 1.2 -fold after 3 h CA treatment in MOLM-14 cells. Numbers in red indicate the percentage of CDK8-occupied genes (peak within ± 5 kb of the gene). **c**, **d**, RNA pol II ChIP-seq metagene profile plots of unchanged genes (black), SE-associated genes (yellow), CA-upregulated genes with vehicle treatment (no CA; red), and CA-upregulated genes with 6 h CA treatment (with CA; blue). **e**, **f**, Cumulative distribution plot of RNA pol II travelling ratio

(TR) after treatment with CA (25 nM, 6 h) or vehicle across genes ≥ 1.2 -fold downregulated by CA after 3 h (1.16-fold, $P = 0.31$, Kolmogorov–Smirnov test) (**e**) and across all genes (1.21-fold, $P < 2.2 \times 10^{-16}$, Kolmogorov–Smirnov test) (**f**). **g**, CA does not significantly change the total amount of RNA or mRNA in MOLM-14 or MV4;11 cells (mean \pm s.e.m., $n = 3$ biological replicates, experiment performed once) after treatment with CA (25 nM, 3 h). **h**, Global levels of RNA pol II pS2 or RNA pol II pS5 do not change after treatment with CA by immunoblot analysis. Flavopiridol (FP) was used at 300 nM as a positive control (experiment performed twice, full scan in Supplementary Fig. 1).



Extended Data Figure 7 | Effects of SE-associated gene expression levels on MOLM-14 AML cell proliferation. **a**, Venn diagram showing overlap between CA-upregulated genes and CD14⁺ master transcription factors. Overlapping genes are listed; SE-associated genes identified by one (purple) or more (red) marks in MOLM-14 are indicated. **b**, GSEA plot showing positive enrichment of CD14⁺ master transcription factors after 3 h CA treatment (MOLM-14 differential expression). **c**, Fold-change in mRNA copies per cell of selected SE-associated genes after 3 h treatment with 100 nM CA, 500 nM I-BET151 or 3 h I-BET151 followed by addition of CA for 3 h (mean \pm s.e.m., $n = 3$ biological replicates, experiment performed twice). **d**, **h**, mRNA expression levels either 1 day (Flag-IRF1, Flag-IRF8) or 3 days (Flag-CDKN1B, Flag-FOSL2, Flag-ETV6) after induction with doxycycline (**d**) or 2 days after siRNA electroporation (**h**) (mean, Poisson error, $n = 15,000$ –20,000).

technical replicates, experiment performed twice) corresponding to Fig. 3f. **e**, Immunoblot showing protein levels of CEBPA 4 days after siRNA electroporation or 1 day after doxycycline-induced expression (experiment performed once) corresponding to Fig. 3f, full scan in Supplementary Fig. 1. **f**, ChIP-seq binding profiles at the *FOSL2* and *ETV6* loci. Red bars denote SEs while grey bars denote regular enhancers. **g**, mRNA levels of indicated genes in MOLM-14 cells expressing Flag-CDK8 (grey) or Flag-CDK8(W105M) (red) after 3 h 25 nM CA treatment (mean \pm s.e.m., $n = 3$ biological replicates, one of two experiments shown). **i**, Heat maps showing BRD4 and CDK8 ChIP-seq on regions depleted of BRD4 > 2-fold after I-BET151 treatment for 6 h before and after drug treatment. **j**, Effect of 3-day treatment with CA, I-BET151 or the combination of CA and I-BET151 on proliferation of MOLM-14 (mean \pm s.e.m., $n = 6$ biological replicates, one of two experiments shown).



Extended Data Figure 8 | CA inhibits AML progression and CDK8 *in vivo* and is well-tolerated at its efficacious dose. **a**, Plasma concentration of CA after single intraperitoneal administration of 1 mg kg^{-1} CA to male CD-1 mice (mean \pm s.e.m., $n = 3$ mice, experiment performed once). **b–g**, MV4;11 disseminated leukaemia study (experiment performed once). **b**, Bioluminescence images with the median bioluminescence for each treatment group on treatment day 1, showing engraftment of MV4;11 leukaemia cells. **c**, 30 days after treatment initiation, the mouse with the highest, lowest, and median day 29 bioluminescence for each treatment group was euthanized and the spleen weight ($P < 0.05$) and percentage of MV4;11 cells (mCherry-positive) in the spleen ($P < 0.03$) and femur bone marrow ($P < 0.02$) were determined ($n = 3$ mice). Dotted purple lines mark the range within 1 s.d. of the mean for the related healthy 8-week-old female NOD-SCID mice, P values determined by one-way ANOVA, each treatment versus vehicle. **d**, Haematoxylin and eosin staining of day-30 lung, spleen and bone marrow samples of the median mice in **c**. Hypercellular alveoli, evidence of leukaemia infiltration, are only observable with vehicle treatment. Spleen sample from the vehicle-treated mouse reveals a large population of cells with a round nucleus and relatively abundant cytoplasm. Similarly, all cells in the vehicle-treated bone marrow have round to oval nuclei and abundant cytoplasm, while normal erythroid or myeloid cells are not observed, suggesting that the spleen and the bone marrow have been dominated by the leukaemia cells. By contrast, the red pulp from the CA-treated mouse spleen shows a heterogeneous population of mature red blood cells, nucleated red blood cells, immature myeloid cells and megakaryocytes. The bone marrow from a CA-treated mouse also exhibits a mixture of erythroid

precursors, myeloid precursors, and megakaryocytes. Scale bars, $250 \mu\text{m}$. **e**, Kaplan–Meier survival analysis ($n = 8$ mice, $P < 0.0001$, log-rank test). **f**, Mean body weight \pm s.e.m., $n = 11$ mice, for study in Fig. 4b. **g**, Complete blood count (CBC) analysis 30 days after first treatment for the mice analysed in **c** ($n = 3$ mice). Dotted purple lines mark the range within 1 s.d. of mean for the related healthy 8-week-old female NOD-SCID mice. **h**, Mean body weight \pm s.e.m., $n = 10$ mice, for study in Fig. 4c (experiment performed once). **i**, Immunoblot of natural killer cell lysate from C57BL/6 mice treated as indicated in Fig. 4d. Each lane represents a distinct mouse sample with 1 = STAT1-pS727, 2 = STAT1, and 3 = β -actin (experiment performed once, full scan in Supplementary Fig. 1). **j–l**, Body weight (**j**), day 15 CBC (**k**), and day 15 blood chemistry (**l**) for healthy CD-1 mice ($n = 3$ mice, experiment performed once) treated with vehicle (20% hydroxypropyl- β -cyclodextrin) or 0.16 mg kg^{-1} CA intraperitoneally once daily for 15 days. **k**, **l**, A/G, albumin/globulin; ALB, albumin (g dl^{-1}); ALK, alkaline phosphatase (U l^{-1}); ALT, alanine aminotransferase (U l^{-1}); AST, aspartate aminotransferase (U l^{-1}); BUN, urea nitrogen (mg dl^{-1}); Ca, total calcium (mg dl^{-1}); CHOL, total cholesterol (mg dl^{-1}); Cl, chloride (mEq l^{-1}); GLOB, globulin (calculated, g dl^{-1}); GLU, glucose (mg dl^{-1}); HCT, haematocrit (%); HGB, haemoglobin (g dl^{-1}); K, potassium (mEq l^{-1}); Na, sodium (mEq l^{-1}); Na/K, sodium/potassium; PHOS, phosphorus (mg dl^{-1}); PLT, platelets ($\times 10^5$ platelets μl^{-1}); RBC, red blood cells ($\times 10^6$ cells per μl); TBIL, total bilirubin (mg dl^{-1}); TP, total protein (g dl^{-1}); TRIG, triglycerides (mg dl^{-1}); WBC, white blood cells ($\times 10^3$ cells μl^{-1}).

Extended Data Table 1 | GI_{50} values for antiproliferative activity of CA and I-BET151

| Cell Line | Malignancy | Mutation | GI_{50} (nM) | |
|-----------|-----------------------|------------------|----------------|----------|
| | | | CA | I-BET151 |
| SKNO-1 | AML | AML1-ETO | 1 | 50 |
| RS4;11 | B-ALL | MLL-AF4 | 3 | 200 |
| SET-2 | AML / MPN | JAK2(V617F) | 4 | 245 |
| MOLM-14 | AML | MLL-AF9 | 5 | 18 |
| MV4;11 | AML | MLL-AF4 | 6 | 20 |
| UKE-1 | AML / MPN | JAK2(V617F) | 7 | |
| MEG-01 | CML / AMKL | BCR-ABL | 9 | 375 |
| TF-1 | AML / Erythroleukemia | EpoR | 350 | 163 |
| HEL | AML / Erythroleukemia | JAK2(V617F) | >1,000 | 200 |
| K562 | CML / Erythroleukemia | BCR-ABL | >1,000 | 750 |
| HCT116 | Colorectal | β -catenin | >1,000 | |

CA and I-BET151 GI_{50} values of human cell lines after 10 and 3 days treatment, respectively, and of HCT116 cells after 7 days ($n = 2$ independent experiments, with three biological replicates each).

Extended Data Table 2 | CA-CDK8-CCNC ternary complex data collection and refinement statistics

| Data collection and refinement statistics (Molecular replacement) | |
|--|--|
| CA-CDK8-CCNC | |
| Data collection | |
| Space group | P 2 ₁ 2 ₁ 2 ₁ |
| Cell dimensions | |
| <i>a</i> , <i>b</i> , <i>c</i> (Å) | 70.5, 71.3, 171.3 |
| α , β , γ (°) | 90.0, 90.0, 90.0 |
| Resolution (Å) | 85.62 (2.40) * |
| <i>R</i> _{sym} | 7.4 (44.8) |
| <i>I</i> / σ <i>I</i> | 10.99 (2.66) |
| Completeness (%) | 94.9 (98.6) |
| Redundancy | 2.8 (2.8) |
| Refinement | |
| Resolution (Å) | 85.62 (2.40) |
| No. reflections | 32875 (8656) |
| <i>R</i> _{work} / <i>R</i> _{free} | 21.7 % / 26.6 % |
| No. atoms | |
| Protein | 5017 |
| Ligand/ion | 50 |
| Water | 104 |
| B-factors | |
| Protein | 32.3 |
| Ligand/ion | 56.3 |
| Water | 47.5 |
| R.m.s deviations | |
| Bond lengths (Å) | 0.009 |
| Bond angles (°) | 1.13 |

*Highest resolution shell is shown in parenthesis.

Crystal structure of human glycine receptor- $\alpha 3$ bound to antagonist strychnine

Xin Huang¹, Hao Chen², Klaus Michelsen¹, Stephen Schneider³ & Paul L. Shaffer¹

Neurotransmitter-gated ion channels of the Cys-loop receptor family are essential mediators of fast neurotransmission throughout the nervous system and are implicated in many neurological disorders. Available X-ray structures of prokaryotic and eukaryotic Cys-loop receptors provide tremendous insights into the binding of agonists, the subsequent opening of the ion channel, and the mechanism of channel activation^{1–8}. Yet the mechanism of inactivation by antagonists remains unknown. Here we present a 3.0 Å X-ray structure of the human glycine receptor- $\alpha 3$ homopentamer in complex with a high affinity, high-specificity antagonist, strychnine. Our structure allows us to explore in detail the molecular recognition of antagonists. Comparisons with previous structures reveal a mechanism for antagonist-induced inactivation of Cys-loop receptors, involving an expansion of the orthosteric binding site in the extracellular domain that is coupled to closure of the ion pore in the transmembrane domain.

Human glycine receptors (GlyRs) are pentameric ligand-gated ion channels (pLGICs) that mediate fast inhibitory synaptic transmission in the spinal cord and brainstem^{9,10}. GlyRs play a key role in motor coordination and the processing of inflammatory pain¹¹. Disruption of the normal function of GlyRs causes hyperekplexia¹², a rare neurological disorder characterized by an exaggerated startle response. GlyRs belong to the large family of Cys-loop receptors, which includes inhibitory anion-selective type A γ -aminobutyric acid receptors (GABA_ARs) together with GlyRs, excitatory cation-selective nicotinic acetylcholine receptors (nAChRs), and serotonin type 3 receptors (5HT₃Rs)¹³. *In vivo*, GlyRs can exist as homopentamers containing only α -subunits or heteropentamers comprising both α - and β -subunits¹⁴. Upon binding of the neurotransmitter glycine to the extracellular domain (ECD), GlyRs undergo conformational changes that allow the transmembrane domain (TMD) to selectively open to permeant anions such as chloride.

X-ray structures of the soluble acetylcholine-binding protein (AChBP), a homologue of the ECD of nAChR, clearly show that the orthosteric binding site is at the subunit interfaces and provide great insight into the binding modes of many agonists and antagonists^{15–18}. Crystal structures of two bacterial pLGIC homologues (ELIC and GLIC)^{1–4} and several eukaryotic pLGICs^{5–8} have recently been solved. The structures of apo⁵ and agonist-bound⁸ (in closed and open conformations respectively) *Caenorhabditis elegans* glutamate-gated chloride channel (GluCl), the agonist-bound human GABA_AR⁶, the nanobody-bound mouse 5HT₃R⁷, and the chimaeric GLIC(ECD)-GlyR $\alpha 1$ (TMD)¹⁹, reveal conformational transitions during activation. Nevertheless, no structural information is available to understand the inactivation mechanism by competitive antagonists. To address this unknown, we solved the crystal structure of a human GlyR, which is a homopentamer of $\alpha 3$ -subunits (GlyR $\alpha 3$) in complex with strychnine. Strychnine is an alkaloid from poisonous plants that causes muscle spasms, convulsions, and eventual death, and remains in use as a rodenticide. Strychnine exerts its lethal effects by antagonizing GlyRs in the central nervous system^{10,20,21}.

To facilitate the formation of well-ordered crystals, we replaced the 76-residue intracellular loop between the transmembrane helix 3 (M3) and helix 4 (M4) with an Ala-Gly-Thr tripeptide and deleted four residues from the carboxy (C) terminus. When solubilized in detergent, GlyR $\alpha 3$ _{cryst} retained the ability to bind strychnine. Surface plasmon resonance (SPR) and isothermal titration calorimetry (ITC) binding studies indicated that GlyR $\alpha 3$ _{cryst} binds to strychnine with a dissociation constant (K_d) of ~ 50 nM, and ITC studies suggested a stoichiometry of five strychnine molecules per GlyR $\alpha 3$ _{cryst} pentamer (Extended Data Fig. 1a, b), in good agreement with published values for the wild-type GlyR $\alpha 3$ (ref. 22). Furthermore, HEK293T cells expressing GlyR $\alpha 3$ _{cryst} displayed glycine-dependent conductance of chloride (Extended Data Fig. 2).

We determined the X-ray structure of GlyR $\alpha 3$ _{cryst} in complex with strychnine at 3.0 Å resolution (Fig. 1, Extended Data Fig. 3 and Extended Data Table 1). Similar to the structures of other pLGICs reported previously, GlyR $\alpha 3$ adopts a cylindrical assembly with a five-fold symmetry and the ion permeation pathway located at the symmetry axis. The ECD comprises an amino (N)-terminal α -helix ($\alpha 1$) followed by a curled β -sandwich with ten β -strands with a second α -helix ($\alpha 2$) between β -strands 3 and 4. Four α -helices (M1–4) after the ECD constitute the TMD, with the 20 helices from the five subunits forming a channel and the five M2 helices lining the channel pore. We observed one glycosylation site, Asn38, and at least one sugar moiety on each subunit. These glycans are solvent exposed and do not interact with other residues, unlike the glycans in the GABA_AR structure⁶.

Strychnine is a very potent and selective antagonist of GlyRs, and binds GlyRs competitively with agonists such as glycine^{10,20,21}. In the co-crystal structure, strychnine binds in a pocket at the interface between adjacent subunits, corresponding to the orthosteric binding site of neurotransmitter agonists in other pLGIC structures (Fig. 2 and Extended Data Fig. 4). The strychnine binding pocket is formed by two loops from the principal or (+) subunit, loop C between strands $\beta 9$ and $\beta 10$ and loop B between strands $\beta 7$ and $\beta 8$, and β -strands from the complementary or (–) subunit, $\beta 1$, $\beta 2$, $\beta 5$, and $\beta 6$. Two Phe residues (63 and 159) form the hydrophobic ‘base’ of the binding pocket for rings V, VI, and VII of strychnine. The ‘flap’ of the binding pocket is composed of residues Tyr202, Thr204, and Phe207 from loop C. Additionally, the backbone carbonyl of Phe159 makes a hydrogen bond with the tertiary amine of strychnine, which is largely protonated at physiological pH ($pK_a = 8.26$). The lactam oxygen is stabilized by the electropositive character of this region of the binding site, in part because of Arg65 (Extended Data Fig. 2f). This binding mode of strychnine in the ‘baseball cap’ shape observed in the co-crystal structure agrees very well with previously published structure–activity relationships of strychnine analogues with GlyRs, which identified the lactam group and the C(21)=C(22) bond as the essential structural features required for strong antagonistic activity towards GlyR $\alpha 1$ and GlyR $\alpha 1\beta^{23,24}$.

The importance of the binding-site residues for strychnine is supported by published mutagenesis data^{25–28}. Mutation of Phe63 to alanine results in over 250-fold loss of affinity and mutation of Phe207 to

¹Department of Molecular Structure and Characterization, Amgen Inc., 360 Binney Street, Cambridge, Massachusetts 02142, USA. ²Department of Protein Technologies, Amgen Inc., 360 Binney Street, Cambridge, Massachusetts 02142, USA. ³Department of Neuroscience, Amgen Inc., 360 Binney Street, Cambridge, Massachusetts 02142, USA.

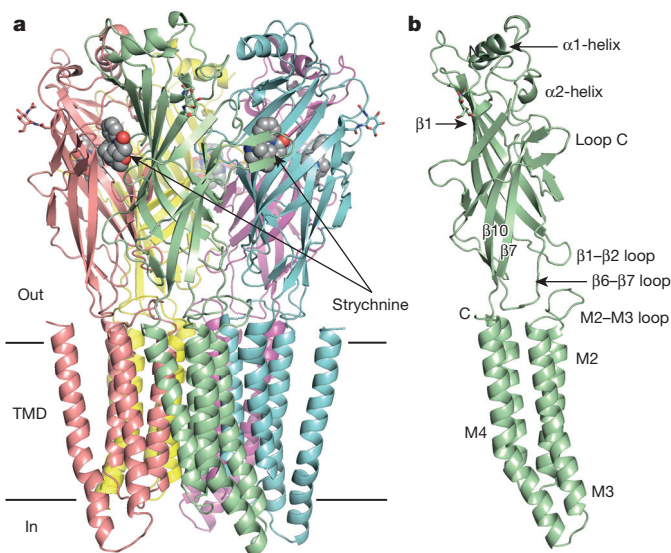


Figure 1 | Architecture of the GlyR α 3 bound to strychnine. **a**, The GlyR α 3–strychnine complex viewed parallel to the plasma membrane. Each subunit is coloured separately. Strychnine is shown as spheres with carbon atoms in grey, nitrogen in blue, and oxygen in red. N-linked glycans are shown as sticks coloured by subunit. **b**, A single subunit of the GlyR α 3 viewed parallel to the membrane but rotated 40° relative to orientation in **a**. Secondary structure elements and important loops are noted.

alanine abolishes strychnine binding. A mutation in rat GlyR α 2 that is equivalent to Gly160Glu in human GlyR α 3 yields striking strychnine insensitivity. In the co-crystal structure, Gly160 C α is in van der Waals contact with the indole ring of strychnine, and addition of the glutamate side chain would prevent strychnine binding in the observed orientation. While the residues critical for strychnine binding are identical among the α - and β -subunits of GlyRs (except that the residue corresponding to Phe207 is a conservatively substituted Tyr in the β -subunit), some of them are not conserved in other pLGICs such as nAChR and 5-HT3R (Extended Data Fig. 5). This explains why strychnine is a competitive antagonist specific to GlyRs.

Superposition of the strychnine-binding site in our co-crystal structure with the agonist-binding sites in the structures of other pLGICs revealed one clear difference (Extended Data Fig. 6). The orthosteric binding site is larger and loop C adopts an open conformation in the strychnine-bound state, reminiscent of antagonist-bound AChBP structures¹⁷. In contrast, the orthosteric site is smaller and loop C adopts a closed conformation in the agonist-bound state of GluCl⁵ and GABA_AR⁶, capping the binding site.

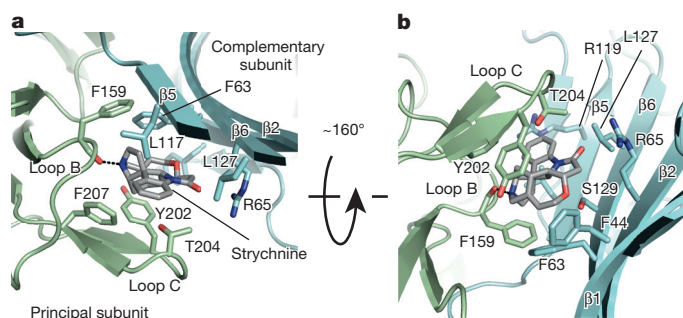


Figure 2 | Orthosteric binding site occupied by the antagonist strychnine. **a**, **b**, Two views of the strychnine-binding site. The view in **a** is from above the receptor binding down, perpendicular to the plasma membrane. Strychnine is shown as grey sticks. The principal subunit is coloured in pale green and the complementary subunit is coloured in cyan. Important residues and secondary structure elements are noted. Black dashed line indicates hydrogen bond from the backbone carbonyl oxygen of Phe159 to the tertiary amine of strychnine.

The ion channel pore of GlyR α 3 is lined by the transmembrane helix M2. There are several constrictions along the length of the GlyR α 3 pore, down to the narrowest 1.4 Å caused by the side chain of Leu261 (Leu 9') in the mid-point of the channel. Since the radius of a dehydrated chloride ion is 1.8 Å, the ion channel of strychnine-bound GlyR α 3 is consistent with a closed, non-conducting state. Leu261 (Leu9') thus forms the shut gate of the ion channel. Leu261 is highly conserved and mutation of the equivalent Leu285 in GlyR β to Arg has been linked to hyperekplexia (Extended Data Fig. 5)¹². Another highly conserved residue Pro250 (Pro-2') of M2 occupies the cytoplasmic end of the ion channel. Pro250 is critical for ion selectivity and mutation Pro250Thr in GlyR α 1 has been linked to hyperekplexia¹². Other hyperekplexia mutations in GlyR α 1 (V260M, T265I, Q266H, and S267N) with spontaneous channel activity¹² are also clustered on the pore-lining M2 helix. In the antagonized state of GlyR α 3, all five M2 helices are straight and oriented parallel to the pore axis (Fig. 3a). In contrast, in the apo structure of GluCl⁸ and the agonist-bound structures of GluCl⁵ and GABA_AR⁶, the M2 helices tilt outwards at the pore apex (Fig. 3b, c). The GlyR α 3–strychnine M2 helices are 11.0 Å apart at the apex (Ala272), 11.7 Å at the base, and 13.0 Å at the most constricted point (Leu261), averaging the distance between C α carbons of i and $i + 2$ protein subunits. In comparison, the apo-GluCl pore is 12.6 Å at the apex (Ser265), 10.3 Å at the base (Pro243), and 12.5 Å at the most constricted point (Leu254). This corresponds to an $\sim 4^\circ$ tilt of the M2 helix towards the pore axis in the strychnine-bound state (Fig. 4a)⁸. Transition from the apo- to the agonist-bound state of GluCl involves tilting of M2 helix by $\sim 8^\circ$ away from the pore axis, which relieves the occlusion of the pore by Leu254 (ref. 5).

For an agonist- or antagonist-binding event to affect the state of the channel, the signal must be transduced across the ECD–TMD interface. Besides the covalent connection by the β 10–M1 linker, the ECD–TMD interface of GlyR α 3 also contains hydrophobic contacts between the N-terminal portion of the M2–M3 loop, the β 1– β 2 loop, and the β 6– β 7 ('Cys') loop as well as polar contacts between the C-terminal portion of the M2–M3 loop and the β 6– β 7 loop (Fig. 4b and Extended Data Fig. 7). Pro275 (from the M2–M3 loop) interacts with Leu142 and Phe145 (from the β 6– β 7 loop) through their side chains and the side-chain hydroxyl oxygen of Tyr279 (of the M2–M3 loop) is hydrogen-bonded to the main-chain amino nitrogen of Leu142 (from the β 6– β 7 loop). The importance of the ECD–TMD interface is highlighted by the hyperekplexia mutations in GlyR α 1 clustered around the M2–M3 loop

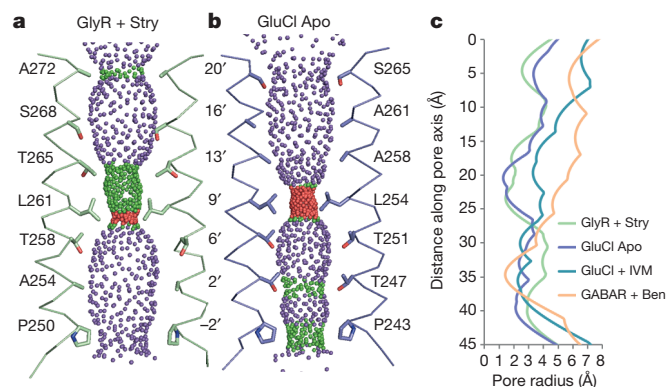


Figure 3 | The ion channel of the GlyR α 3–strychnine complex in a closed state. **a**, Solvent contours of the transmembrane pore of the strychnine-bound GlyR α 3 pore showing the M2 helices of subunits A and C. Side chains of pore-lining residues are shown. Numbering is according to the protein sequence and position in the M2 helix. Small purple, green, and red spheres define a radius of >2.8 Å, 1.4–2.8 Å, and <1.4 Å, respectively. **b**, Contours of the apo-GluCl pore, similar to **a**. **c**, Plot of pore radii as a function of distance along the pore axis for strychnine-bound GlyR α 3, apo- and ivermectin-bound GluCl α , and benzamidine-bound GABA_AR- β 3; Stry, strychnine; IVM, ivermectin; Ben, benzamidine.

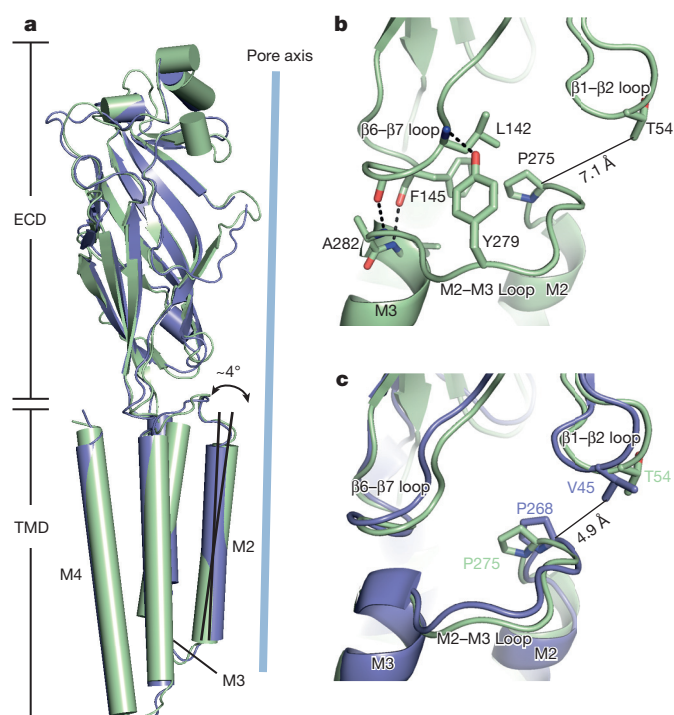


Figure 4 | Structural coupling at the ECD-TMD interface. **a**, Superposition of the ECDs of strychnine-bound GlyR α 3 (green) and apo-GluCl (blue) illustrates an $\sim 4^\circ$ tilt of the M2 pore-lining helix towards the pore axis. **b**, Illustration of key residues forming hydrogen bond and hydrophobic interactions that connect the ECD and TMD. Hydrogen bonds are denoted by dashed black lines. The average distance from Pro275 C α to Thr54 C γ is noted. **c**, Superposition of the entire pentameric complex of strychnine-bound GlyR α 3 (green) and apo-GluCl (blue) reveals the position of the M2-M3 loop relative to the β 1- β 2 loop. The average distance from Pro268 C α to Val45 C γ in the apo-GluCl pentamer is noted.

(R271Q/L/P, K276E/Q, and Y279C/S) as they lead to reductions in glycine sensitivity and maximum probability of channel opening (Extended Data Fig. 5)¹².

All-atom molecular dynamics simulations of GluCl predicted that the unbinding of agonist at the orthosteric site and the opening of the orthosteric site lead to repositioning of the β 1- β 2 loop and inward displacement of the M2-M3 loop towards the pore, which is then coupled to the untilting of the M2 helix and the closing of the pore²⁹. This gating mechanism was subsequently validated by the apo-GluCl structure⁸ where the pore is closed, M2 helix is untilted from the pore by $\sim 8^\circ$, and the M2-M3 loop shifts by more than 6 Å away from the channel pore, as visualized by the movement of Pro268 of the M2-M3 loop passing beneath Val45 on the β 1- β 2 loop. The M2-M3 loop of our strychnine-bound GlyR α 3 is much closer to the pore centre than in apo-GluCl structure (Fig. 4b). Pro275 of the M2-M3 loop is 7.1 Å from Thr54 of the β 1- β 2 loop in the strychnine-bound GlyR α 3 structure while the equivalent distance is 4.9 Å between Pro268 and Val45 in the apo-GluCl structure (Fig. 4c), suggesting that the M2-M3 loop in GlyR α 3 is pulled even more towards the pore than in apo-GluCl. Similar movement of the M2-M3 loop and the β 1- β 2 loop upon strychnine binding has also been observed in molecular dynamics simulations of a homology model of GlyR α 1 (ref. 30). This could lead to more energetic stabilization of the untilting of the M2 helix and the closing of the pore. Therefore, we hypothesize that the binding of the antagonist strychnine to GlyR α 3 induces the opening of loop C and the orthosteric binding pocket, which further facilitates the inward displacement of the M2-M3 loop and the tilting of the M2 helix towards the pore, ultimately leading to closing of the channel pore (Fig. 5).

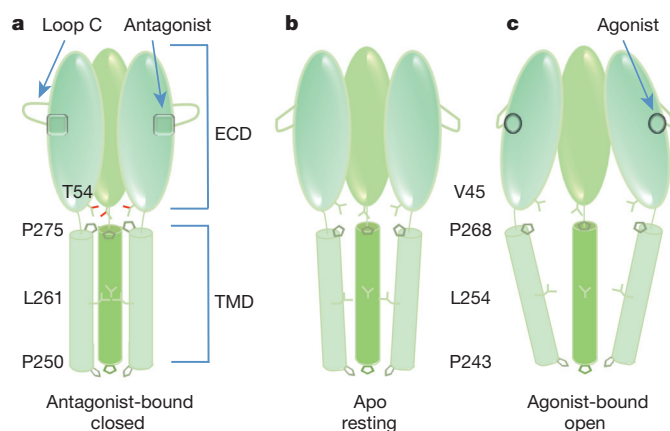


Figure 5 | Conformational changes in pLGICs. **a**, Schematic illustration of the conformation of the strychnine-bound closed GlyR α 3. **b**, Schematic illustration of the conformation of the apo resting GluCl. **c**, Schematic illustration of the conformation of the glutamate/ivermectin-bound open GluCl.

In summary, we present an X-ray structure of a GlyR, the human GlyR α 3 homopentamer, co-crystallized with the antagonist strychnine. Strychnine is bound in a larger orthosteric pocket and loop C of GlyR α 3 adopts an open conformation. The M2-M3 loop is displaced inwards, the pore-lining M2 helix tilts in a direction opposite to that observed in the active conformation of related pLGICs, and the pore is shut. Our study represents the first crystallographic analysis of a pLGIC in the inactive state induced by a competitive antagonist. These results shed new light on the conformational transitions upon antagonist binding and provide a rational basis for understanding human hyperekplexia mutations and the specificity of strychnine for GlyRs.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 1 April; accepted 23 July 2015.

Published online 28 September 2015.

- Hilf, R. J. & Dutzler, R. X-ray structure of a prokaryotic pentameric ligand-gated ion channel. *Nature* **452**, 375–379 (2008).
- Hilf, R. J. & Dutzler, R. Structure of a potentially open state of a proton-activated pentameric ligand-gated ion channel. *Nature* **457**, 115–118 (2009).
- Bocquet, N. *et al.* X-ray structure of a pentameric ligand-gated ion channel in an apparently open conformation. *Nature* **457**, 111–114 (2009).
- Sauguet, L. *et al.* Crystal structures of a pentameric ligand-gated ion channel provide a mechanism for activation. *Proc. Natl Acad. Sci. USA* **111**, 966–971 (2014).
- Hibbs, R. E. & Gouaux, E. Principles of activation and permeation in an anion-selective Cys-loop receptor. *Nature* **474**, 54–60 (2011).
- Miller, P. S. & Aricescu, A. R. Crystal structure of a human GABA_A receptor. *Nature* **512**, 270–275 (2014).
- Hassaine, G. *et al.* X-ray structure of the mouse serotonin 5-HT₃ receptor. *Nature* **512**, 276–281 (2014).
- Althoff, T., Hibbs, R. E., Banerjee, S. & Gouaux, E. X-ray structures of GluCl in apo states reveal a gating mechanism of Cys-loop receptors. *Nature* **512**, 333–337 (2014).
- Legendre, P. The glycinergic inhibitory synapse. *Cell. Mol. Life Sci.* **58**, 760–793 (2001).
- Lynch, J. W. Molecular structure and function of the glycine receptor chloride channel. *Physiol. Rev.* **84**, 1051–1095 (2004).
- Lynch, J. W. & Callister, R. J. Glycine receptors: a new therapeutic target in pain pathways. *Curr. Opin. Investig. Drugs* **7**, 48–53 (2006).
- Bode, A. & Lynch, J. W. The impact of human hyperekplexia mutations on glycine receptor structure and function. *Mol. Brain* **7**, 2 (2014).
- Schofield, P. R. *et al.* Sequence and functional expression of the GABA_A receptor shows a ligand-gated receptor super-family. *Nature* **328**, 221–227 (1987).
- Langosch, D., Thomas, L. & Betz, H. Conserved quaternary structure of ligand-gated ion channels: the postsynaptic glycine receptor is a pentamer. *Proc. Natl Acad. Sci. USA* **85**, 7394–7398 (1988).
- Brejck, K. *et al.* Crystal structure of an ACh-binding protein reveals the ligand-binding domain of nicotinic receptors. *Nature* **411**, 269–276 (2001).

16. Celie, P. H. *et al.* Nicotine and carbamylcholine binding to nicotinic acetylcholine receptors as studied in AChBP crystal structures. *Neuron* **41**, 907–914 (2004).
17. Hansen, S. B. *et al.* Structures of *Aplysia* AChBP complexes with nicotinic agonists and antagonists reveal distinctive binding interfaces and conformations. *EMBO J.* **24**, 3635–3646 (2005).
18. Billen, B. *et al.* Molecular actions of smoking cessation drugs at $\alpha 4\beta 2$ nicotinic receptors defined in crystal structures of a homologous binding protein. *Proc. Natl Acad. Sci. USA* **109**, 9173–9178 (2012).
19. Moraga-Cid, G. *et al.* Allosteric and hyperekplexic mutant phenotypes investigated on an α_1 glycine receptor transmembrane structure. *Proc. Natl Acad. Sci. USA* **112**, 2865–2870 (2015).
20. Rajendra, S., Lynch, J. W. & Schofield, P. R. The glycine receptor. *Pharmacol. Ther.* **73**, 121–146 (1997).
21. Laube, B., Maksay, G., Schemm, R. & Betz, H. Modulation of glycine receptor function: a novel approach for therapeutic intervention at inhibitory synapses? *Trends Pharmacol. Sci.* **23**, 519–527 (2002).
22. Grenningloh, G. *et al.* Alpha subunit variants of the human glycine receptor: primary structures, functional expression and chromosomal localization of the corresponding genes. *EMBO J.* **9**, 771–776 (1990).
23. Jensen, A. A., Gharagozloo, P., Birdsall, N. J. & Zlotos, D. P. Pharmacological characterisation of strychnine and brucine analogues at glycine and $\alpha 7$ nicotinic acetylcholine receptors. *Eur. J. Pharmacol.* **539**, 27–33 (2006).
24. Mohsen, A. M., Heller, E., Holzgrabe, U., Jensen, A. A. & Zlotos, D. P. Structure–activity relationships of strychnine analogs at glycine receptors. *Chem. Biodivers.* **11**, 1256–1262 (2014).
25. Grudzinska, J. *et al.* The β subunit determines the ligand binding properties of synaptic glycine receptors. *Neuron* **45**, 727–739 (2005).
26. Brams, M. *et al.* A structural and mutagenic blueprint for molecular recognition of strychnine and *d*-tubocurarine by different Cys-loop receptors. *PLoS Biol.* **9**, e1001034 (2011).
27. Becker, C. M., Hoch, W. & Betz, H. Glycine receptor heterogeneity in rat spinal cord during postnatal development. *EMBO J.* **7**, 3717–3726 (1988).
28. Vandenberg, R. J., French, C. R., Barry, P. H., Shine, J. & Schofield, P. R. Antagonism of ligand-gated ion channel receptors: two domains of the glycine receptor α subunit form the strychnine-binding site. *Proc. Natl Acad. Sci. USA* **89**, 1765–1769 (1992).
29. Calimet, N. *et al.* A gating mechanism of pentameric ligand-gated ion channels. *Proc. Natl Acad. Sci. USA* **110**, E3987–E3996 (2013).
30. Yu, R. *et al.* Agonist and antagonist binding in human glycine receptors. *Biochemistry* **53**, 6041–6051 (2014).

Acknowledgements We thank G. Ranieri and R. Walter at Shamrock Structures and the staff at beamlines 08-ID at the Canadian Light Source and 22-ID at the Advanced Photon Source for data collection. We are grateful to Z. Wang and J. Gingras for reviewing the manuscript.

Author Contributions The authors have jointly contributed to project design, data analysis and manuscript preparation. P.L.S. performed initial construct design and purification experiments, structure solution, model building, and structural analysis; X.H. performed protein purifications, crystallization, model building, and structural analysis; H.C. performed cloning and expression experiments; K.M. performed SPR and ITC binding studies; S.S. performed functional testing; P.L.S. and X.H. wrote the manuscript with help from K.M., S.S., and H.C.

Author Information Atomic coordinates and structure factors for the GlyR $\alpha 3$ –strychnine complex have been deposited in the Protein Data Bank (PDB) under accession number 5CFB. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to X.H. (hxin@amgen.com) or P.L.S. (pschaffer@amgen.com).

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

GlyR α 3 construct, expression, and purification. The GlyR α 3_{cryst} protein construct was derived by removal of the M3–M4 loop on the basis of alignments with bacterial channels and the *C. elegans* GluCl α crystallization construct and replaced with a tri-peptide linker (AGT). The C terminus was truncated by four residues on the basis of sequence alignments with various mammalian GlyR α 3 sequences. To facilitate purification, a Strep II affinity tag (WSHPQFEK) was added to the C terminus. In total, GlyR α 3_{cryst} corresponds to the polypeptide sequence of human GlyR α 3 (Uniprot O75311) 1–460 (Δ 343–418::AGT)–Strep. The recombinant baculovirus was generated with the Bac-to-Bac system (Life Technologies). Expression was done by baculovirus transduction of Sf9 insect cells grown in sfx medium (Hyclone) at 27 °C for 72 h. Cells were harvested by centrifugation at 2,000g and disrupted in an Microfluidizer in a buffer containing 50 mM Tris pH 8.0, 150 mM NaCl, and 1% protease inhibitors cocktail (Sigma). The homogenate was clarified by centrifugation at 10,000g and crude membranes were collected by centrifugation at 125,000g. The membrane were mechanically homogenized and solubilized in 0.2 g DDM per gram of membranes in 20 mM Tris pH 8.0, 150 mM NaCl, 0.5% protease inhibitors cocktail. Solubilized membranes were centrifuged at 125,000g and supernatant was bound to Strep affinity resin (IBA), washed with 20 mM Tris pH 8.0, 150 mM NaCl, 1 mM DDM, and eluted with 20 mM Tris pH 8.0, 150 mM NaCl, 1 mM DDM, and 5 mM desthiobiotin. Eluted fractions containing GlyR α 3_{cryst} were pooled together, concentrated, and further purified by gel filtration in 20 mM Tris pH 8.0, 150 mM NaCl, and 1 mM DDM. All purification steps were performed at 4 °C.

Crystallization and data collection. Purified GlyR α 3 were concentrated to \sim 3 mg ml^{−1} and incubated with 0.2 mM strychnine at 4 °C for 30 min before crystallization. Crystals of GlyR α 3–strychnine were obtained in hanging drop at 4 °C by mixing 0.5 μ l of the GlyR α 3–strychnine complex with 0.5 μ l of the crystallization buffer containing 25 mM sodium citrate pH 4.0, 100 mM KCl, 200 mM MgCl₂, 30–33% PEG400. Crystals take about 1 month to grow and were frozen directly from the crystallization drops in liquid nitrogen for data collection. Diffraction data were collected at beamline 22-ID at the Advanced Photon Source, Argonne National Laboratory. Diffraction data were indexed, integrated, and scaled using HKL2000 software³¹. Despite screening many crystals at the synchrotron, the highest resolution data set obtained from a single crystal was 3.5 Å owing to their relatively small size and rapid degradation upon exposure to high-intensity X-ray beams. The final complete 3.0 Å data set was assembled by merging data from 15 separate crystals with 10–30° of data collected from each crystal to limit radiation damage (Extended Data Table 1).

Structure determination. Initial phases for the structure were generated by molecular replacement with Phaser³² using the pentameric apo-GluCl structure as a search model (PDB accession number 4TNV)⁸. A clear solution was obtained with one pentameric assembly of GlyR α 3 in the asymmetric unit. Electron density maps were improved by fivefold non-crystallographic symmetry averaging. Initial maps were of sufficient quality to build strychnine molecules and GlyR α 3 side chains where they differed from those of GluCl. In agreement with established conventions, the residue numbering scheme reassigns residue Ala34 of the Uniprot sequence as Ala1 because it is the first amino acid in the mature polypeptide following removal of the secretion signal sequence. Additionally, residue numbering in the GlyR α 3_{cryst} model is continuous, meaning residues following the deletion of the intracellular domain between M3 and M4 do not retain their numbering from the wild-type protein. The restraint parameters for strychnine were generated by PRODRG³³. Iterative rounds of restrained refinement in Refmac5³⁴ and manual rebuilding in Coot³⁵ were used to improve the model. In final rounds of refinement, fivefold non-crystallographic symmetry restraints were removed and ten TLS parameters added, one each for the ECD and the TMD of the five protomers in the structure. Model quality was assessed using Molprobity³⁶. The final model consists of one GlyR α 3 pentamer including residues 9–347, N-linked glycosylation at Asn38, and five strychnine molecules (Extended Data Table 1). Electrostatic surface potential calculations were performed using the APBS³⁷ Tool plug-in in PyMOL and pore dimensions were analysed with HOLE³⁸. Images were made using PyMOL³⁹. Sequence alignment was performed in ClustalW⁴⁰.

ITC and SPR binding experiments. ITC experiments were performed on an ITC200 instrument (GE Healthcare). Protein concentration was determined by absorbance at 280 nm using a molar extinction coefficient per cm of 66,600. For titration experiments, GlyR α 3 was diluted to 7 μ M in phosphate buffered saline (PBS) pH 7.4 and placed in the ITC cell. Freshly prepared strychnine (10 mM stock in dimethylsulfoxide (DMSO)) was diluted to 100 μ M in PBS pH 7.4 and placed in the syringe. Final DMSO concentration in the cell and syringe was 1% (v/v).

Strychnine titration into buffer was also performed to ensure minimal heat of dilution. Titrations were performed at 25 °C using 2 μ l injections (4 s duration, 180 s spacing, and 5 s filter period). Reference power was set to 10 μ cal s^{−1} and stirring speed to 1,000 r.p.m. The raw data were baseline corrected and integrated using Origin 7.0. All thermodynamic parameters were obtained from the fitting of the heat data assuming 1:1 interaction model. Binding experiments were performed twice ($n = 2$) to calculate a standard deviation.

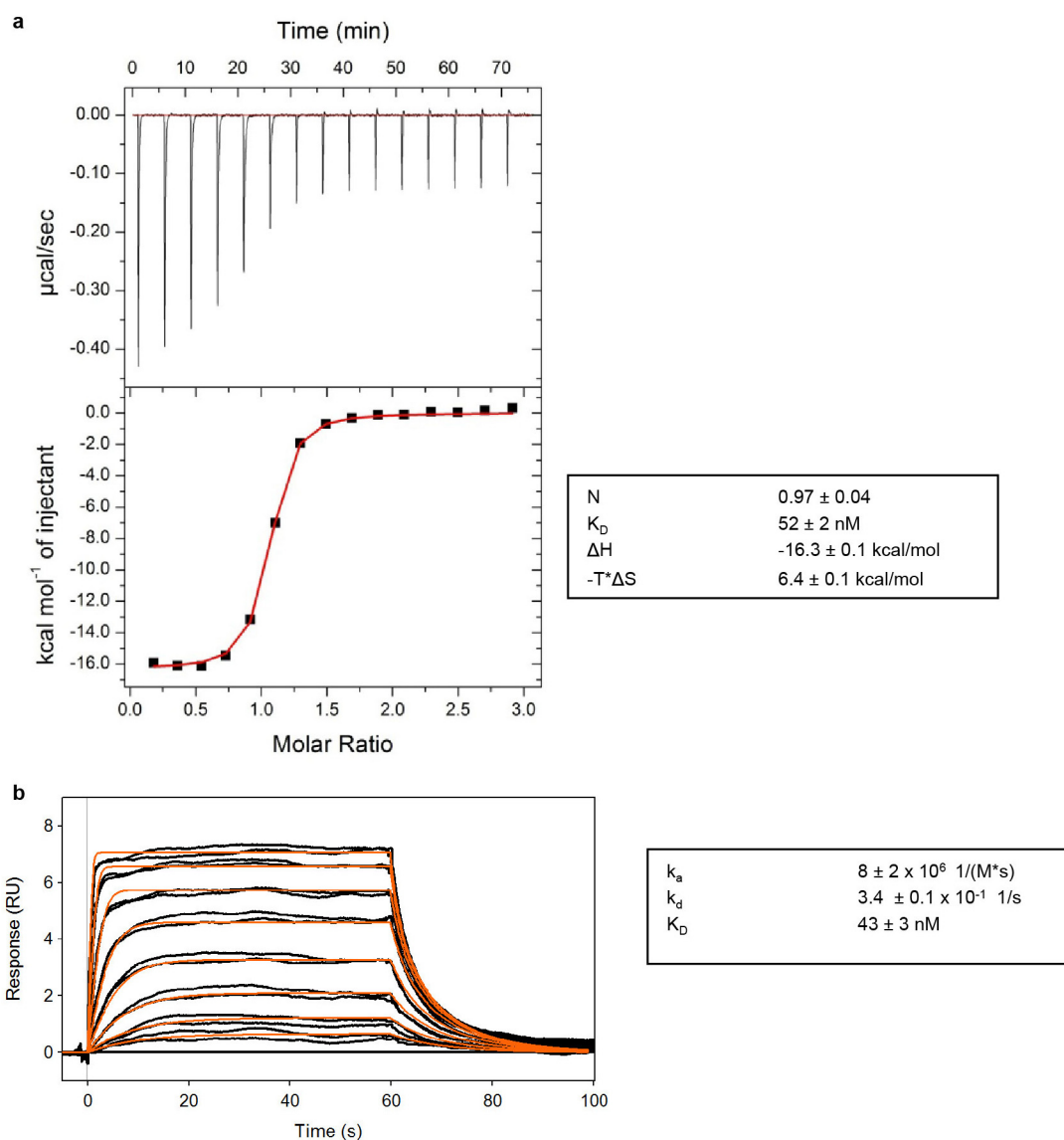
SPR measurements were performed on a Biacore T200 (GE Healthcare) at 25 °C using PBS pH 7.4 as running buffer. L1 chips were pre-conditioned with three 30 s injections of 20 mM CHAPS. GlyR α 3 was diluted to 300 μ g ml^{−1} in PBS pH 7.4 containing 1 mM DDM and 0.1 mg ml^{−1} POPG. GlyR α 3 was passed over L1 chip and captured onto the biosensor at a density of 4,000 response units. In the reference cell, solution with lipid (that is, no GlyR α 3) was injected. DMSO at 1% (v/v) was added to the running buffer and strychnine injected at various concentrations (top concentration 2 μ M, twofold dilution series, in duplicate) at a flow rate of 90 μ g min^{−1}. The association was set to 1 min followed by 2 min dissociation. The raw data were processed using Scrubber2 software (BioLogic Software) and the data kinetically fitted to a 1:1 binding model which included a mass transfer limitation term. Binding experiments were performed three times ($n = 3$) to calculate a standard deviation.

Functional testing in FLIPR assay⁴¹. HEK293T cells were cultured in Cell Culture Media (MEM supplemented with 10% v/v qualified heat-inactivated FBS, 100 units penicillin, 100 units streptomycin, 29.2 mg ml^{−1} of L-glutamine) under standard cell culture conditions of 37 °C, 5% v/v CO₂, and 95% humidity. Cells were grown in T 225 cm² culture flasks to a density of approximately 8×10^7 cells and harvested after about 4 days by briefly washing with DPBS followed by addition of Cell Dissociation Reagent for 2 min. The concentration of cells in suspension was adjusted to 6.40×10^5 cells per millilitre in Cell Plating Media (MEM with 10% dialysed FBS, 100 units penicillin, 100 units streptomycin, 0.29 mg ml^{−1} of L-glutamine and 10 mM HEPES pH 7.4) and transduced with either wild-type GlyR α 3 (MOI = 1) or GlyR α 3_{cryst} (MOI = 5) baculovirus containing a BacMam vector with a CMV promoter. Using a Multidrop Combi, 25 μ l of cell suspension was dispensed into Corning CellBIND 384-well ViewPlates. Transduced cell culture plates were then incubated at 37 °C overnight under the standard cell-culture conditions described above. The next day (about 18–24 h after plating), 5 μ l of 6 \times Membrane Potential blue dye for monitoring changes in membrane potential was dispensed into each cell culture plate using a Thermo Multidrop Combi (prepared in assay buffer at 6 \times the manufacturer's recommended final concentration). The cell plates were then incubated at 37 °C for 30 min and allowed to equilibrate to room temperature (25 °C) for an additional 30 min.

Glycine dose–response plates were prepared in assay buffer (10 mM HEPES, 60 mM NaCl, 5 mM KCl, 2 mM MgCl₂, 1 mM CaCl₂, 10 mM D-glucose, 160 mM D-mannitol and 2 M KOH solution to adjust pH to 7.4) supplemented with 2% v/v DMSO using a 1:2 stepwise dilution series in standard 384-well polypropylene plates. The membrane potential assay is performed on FLIPR Tetra, which transfers 10 μ l from the 4 \times glycine dose–response plate and adds it to the 30 μ l volume in each well of the cell plate containing Membrane Potential blue dye. Fluorescence emission (510–545 nm/565–625 nm excitation/emission filter set, excitation intensity 40%, camera gain 50, and an exposure time of 0.4 s) is measured in real-time to detect changes in membrane potential. The net membrane potential of the cells changes upon activation of GlyR α 3, which results in the increased flux of chloride ions out of the cell down a concentration gradient and a robust increase in fluorescence signal. FLIPR kinetic traces were processed using an area under the curve relative to baseline algorithm, where the baseline was the first 10 s of the measurement before addition of glycine to the cell plate. All measurements proceeded for 120 s post-addition. The processed data were subsequently normalized to the maximum achievable glycine response (at 2 mM glycine). Normalized data were then plotted against log[glycine] and data were fitted to a nonlinear regression four-parameter Hill fit to determine the half-maximum effective concentration/half-maximum inhibitory concentration (EC₅₀/IC₅₀) from the resulting sigmoidal curve. All curve fitting was performed with GraphPad Prism 6 software.

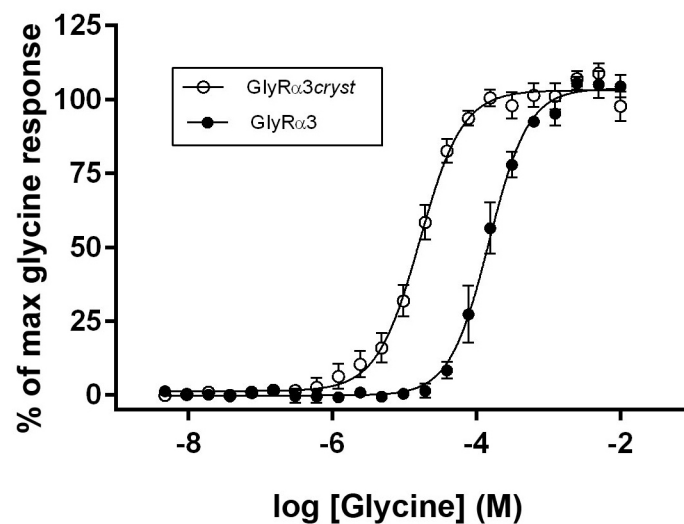
- Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
- McCoy, A. J. Solving structures of protein complexes by molecular replacement with Phaser. *Acta Crystallogr. D* **63**, 32–41 (2007).
- Schüttelkopf, A. W. & van Aalten, D. M. PRODRG: a tool for high-throughput crystallography of protein-ligand complexes. *Acta Crystallogr. D* **60**, 1355–1363 (2004).
- Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D* **53**, 240–255 (1997).
- Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
- Davis, I. W. *et al.* MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.* **35**, W375–W383 (2007).

37. Baker, N. A., Sept, D., Joseph, S., Holst, M. J. & McCammon, J. A. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl Acad. Sci. USA* **98**, 10037–10041 (2001).
38. Smart, O. S., Neduvellil, J. G., Wang, X., Wallace, B. A. & Sansom, M. S. HOLE: a program for the analysis of the pore dimensions of ion channel structural models. *J. Mol. Graph.* **14**, 354–360, 376 (1996).
39. DeLano, W. L. The PyMOL molecular graphics system (DeLano Scientific, 2002).
40. Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948 (2007).
41. Jensen, A. A. & Kristiansen, U. Functional characterisation of the human $\alpha 1$ glycine receptor in a fluorescence-based membrane potential assay. *Biochem. Pharmacol.* **61**, 1789–1799 (2004).



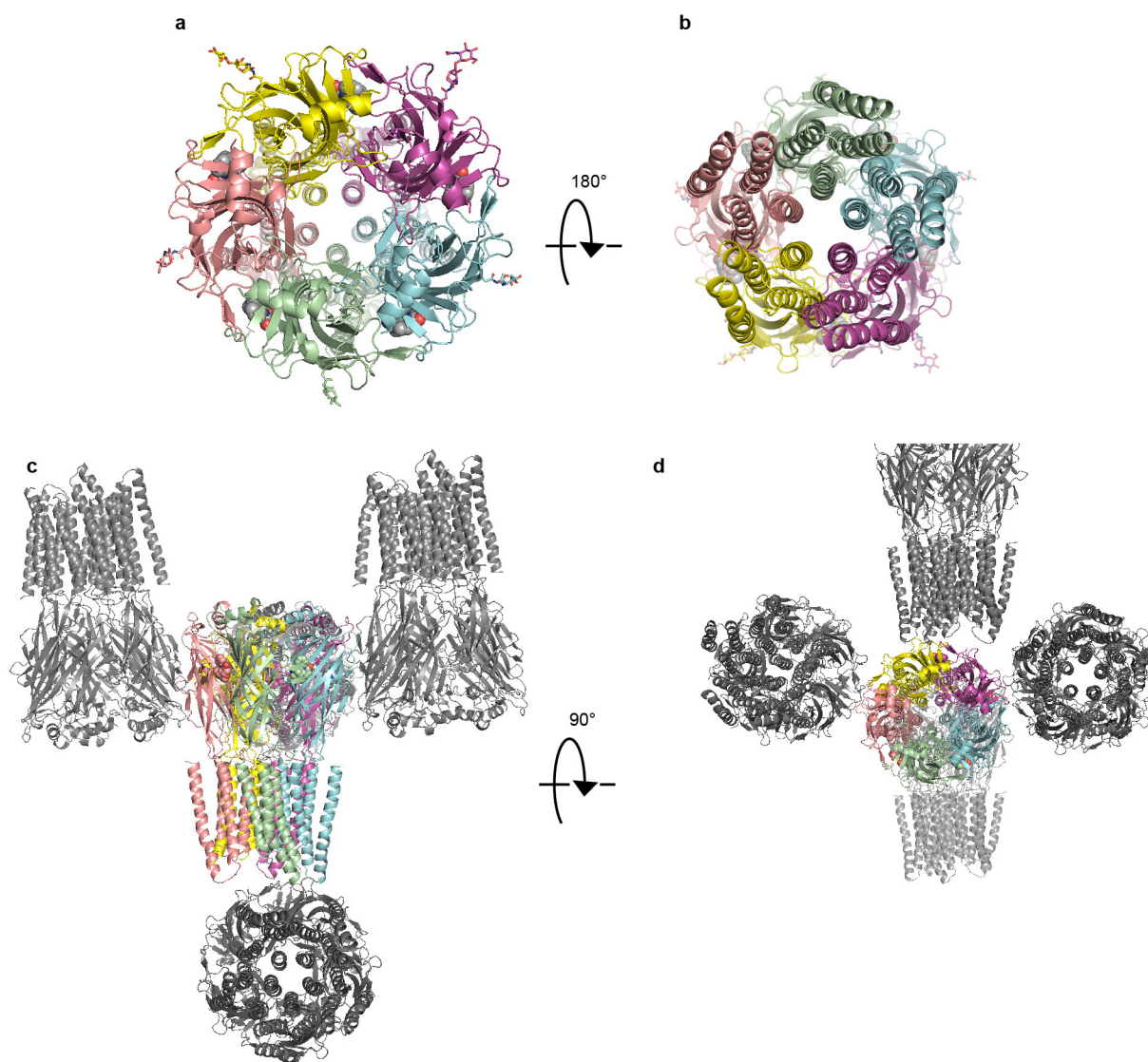
Extended Data Figure 1 | Strychnine binding to GlyR α 3_{cryst} in detergent micelles. **a**, Binding thermodynamics and stoichiometry of strychnine to GlyR α 3_{cryst} analysed by ITC. The individual peaks from titrations are integrated and presented in a Wiseman plot. An appropriate binding model is

chosen and the isotherm is then fitted to yield the binding enthalpy ΔH , the dissociation constant K_D , and the stoichiometry n . **b**, Binding kinetics of strychnine to GlyR α 3_{cryst} measured by SPR spectroscopy.



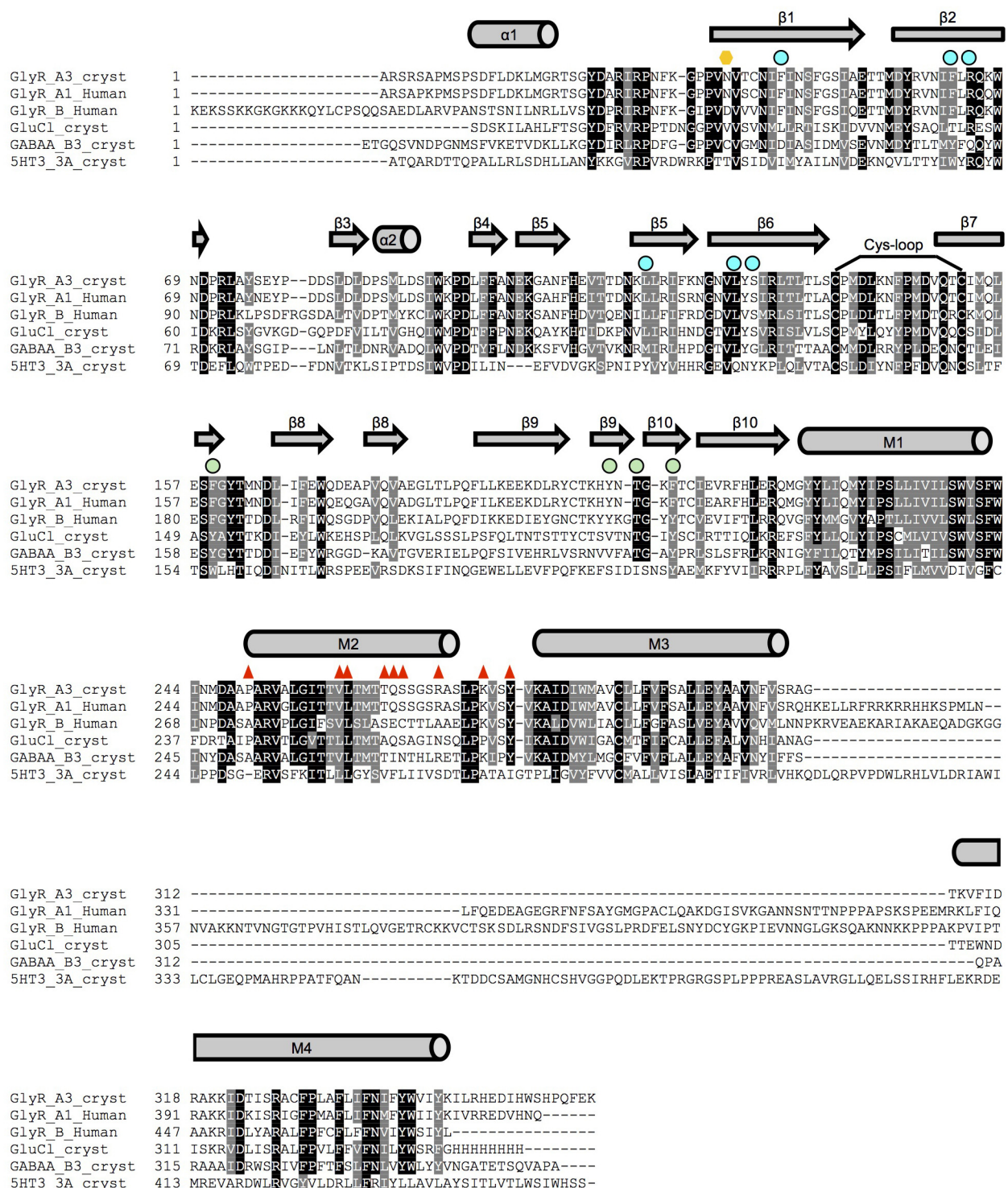
Extended Data Figure 2 | Glycine dose–response to GlyRα3_{cryst} in HEK293T cells. Glycine dose–response curve for BacMam baculovirus-transduced HEK293T cells expressing either wild-type human GlyRα3 (GlyRα3, filled circles) or GlyRα3_{cryst} (open circles) measured by membrane potential dye assay (see Methods for details). Each data point represents a value

of $n = 4–6$ and normalized to a maximum glycine response observed at 2 mM glycine concentration. Glycine EC₅₀ for GlyRα3 was calculated at $150 \pm 10.6 \mu\text{M}$ ($n = 4$, 95% confidence interval) and GlyRα3_{cryst} at $16.4 \pm 1.2 \mu\text{M}$ ($n = 5$, 95% confidence interval).



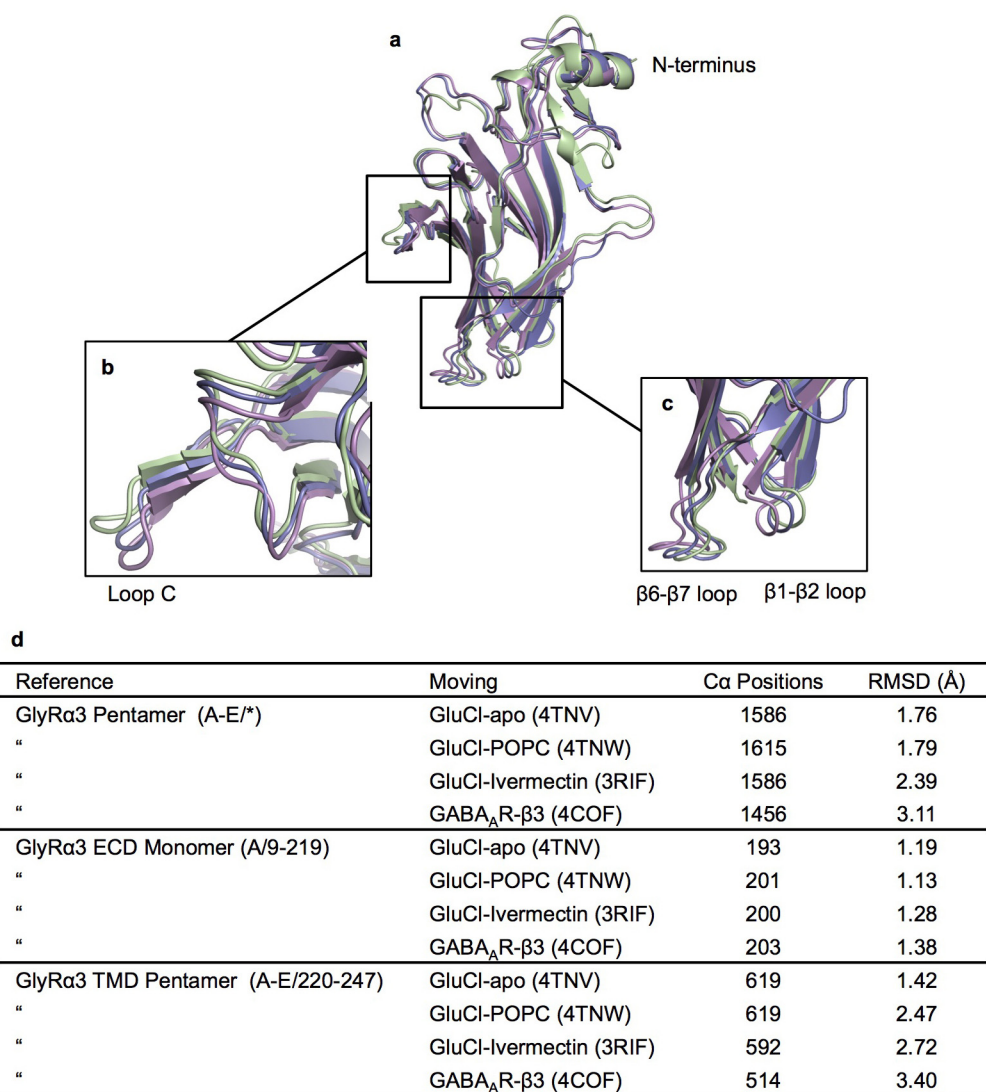
Extended Data Figure 3 | Architecture of the GlyRα3 bound to strychnine and crystallographic packing of GlyRα3_{cryst}. **a**, The GlyRα3–strychnine complex viewed from the extracellular side of the membrane down the pore axis, perpendicular to the membrane. Strychnine is bound at the interface between subunits. **b**, The GlyRα3–strychnine complex viewed from the intracellular side down the pore axis. The M2 helices are shown lining the pore. **c**, **d**, Packing of the GlyRα3–strychnine complex. Receptor in the asymmetric

unit is coloured by subunit, with subunit A in pale green and subunit B in cyan. Strychnine molecules are shown as spheres and are coloured to match their associated principal subunit. Symmetry-related receptors are coloured grey. The interface between subunits A and B is completely exposed to solvent in the crystal lattice, and this interface was used in the making of all figures for this paper to avoid any potential crystal-packing artefacts.

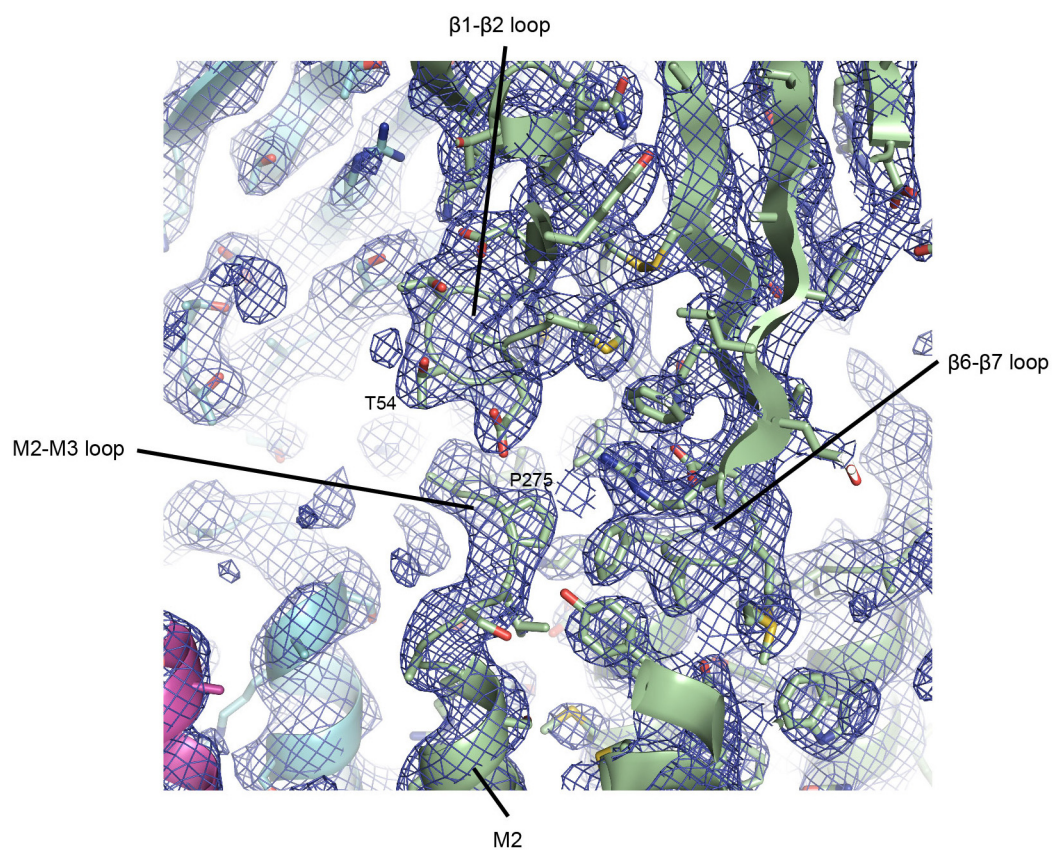


Extended Data Figure 5 | Sequence alignment of GlyR α ₃_{cryst} with representative eukaryotic Cys-loop receptor family members. Residue conservation is indicated by grey and black highlights. Site of N-linked glycosylation of GlyR α ₃_{cryst} is indicated by the orange hexagon. Residues involved in binding of strychnine are indicated by green (the principal subunit) and cyan (the complementary subunit) dots. Red triangles above residues indicate mutations in GlyR α ₁ or GlyR β that cause hyperekplexia. Signal

peptides have been removed from all protein sequences. Secondary structure elements are denoted by cylinders (helices) and arrows (strands) above the alignment. The alignment was generated using ClustalW. Protein sequences are from the following entries: human GlyR α ₁ (Uniprot P23415), human GlyR β (Uniprot 48167), GluCl_{cryst} (PDB accession number 4TNV), GABA_A- β ₃_{cryst} (PDB accession number 4COF), and 5HT_{3A} (PDB accession number 4PIR).



Extended Data Figure 6 | Structural alignment of pLGICs. **a**, Structures of GlyRα₃_{cryst} (green), apo-GluClα (slate), and glutamate- and ivermectin-bound GluCLα (violet) aligned using their ECDs. **b**, Close-up view of the loop C, rotated 45° clockwise along the horizontal axis (left to right) to optimize viewing. **c**, Close-up view of loops β6–β7 and β1–β2. **d**, Table showing regions of GlyRα₃ used for structural alignments, as well as resulting number of overlapped Cα atoms and root mean squared deviation distances. PDB accession numbers are listed for each reference structure.



Extended Data Figure 7 | Representative electron density for the ECD-TMD interface. Final $2F_o - F_c$ electron density map around the region of the ECD-TMD interface is shown, contoured at 1.0σ . Protomer A is shown in pale

green, protomer B in cyan, and protomer C in magenta. Important secondary structure elements and loops are noted.

Extended Data Table 1 | Crystallographic and structure refinement statistics

| | GlyR α 3 _{cryst} + strychnine |
|---|---|
| Data collection | |
| Space group | P2 ₁ 2 ₁ 2 ₁ |
| Cell dimensions | |
| <i>a</i> , <i>b</i> , <i>c</i> (Å) | 140.2, 140.2, 180.1 |
| α , β , γ (°) | 90.0, 90.0, 90.0 |
| Resolution (Å) | 50-3.04 (3.15-3.04) * |
| <i>R</i> _{merge} | 0.209 (1.177) |
| <i>R</i> _{pim} | 0.064 (0.452) |
| <i>I</i> / σ <i>I</i> | 8.5 (1.3) |
| Completeness (%) | 96.9 (77.6) |
| Redundancy | 10.1 (5.8) |
| CC _{1/2} | 0.983 (0.573) |
| Refinement | |
| Resolution (Å) | 50-3.04 |
| No. reflections | 67970 |
| <i>R</i> _{work} / <i>R</i> _{free} | 0.243/0.258 |
| No. atoms | |
| Protein | 13317 |
| Ligand/ion | 223 |
| Water | 0 |
| B-factors | |
| Protein | 124 |
| Ligand/ion | 120 |
| R.m.s deviations | |
| Bond lengths (Å) | 0.006 |
| Bond angles (°) | 1.05 |

This structure was determined from fifteen crystals.

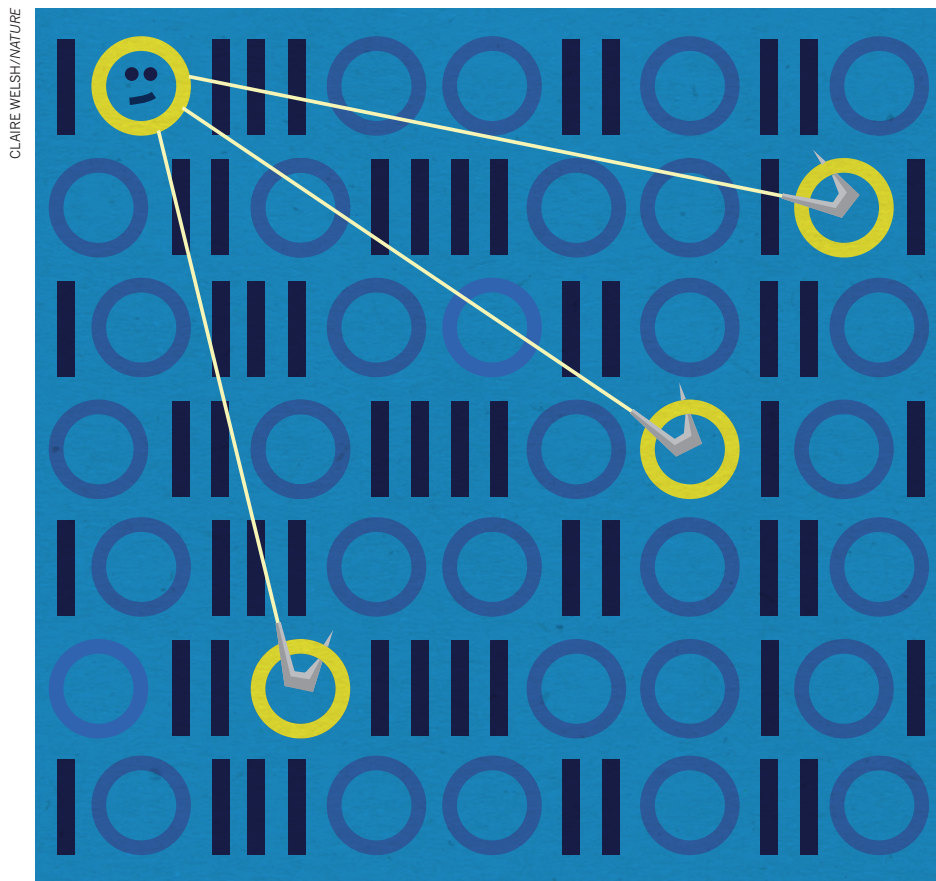
*Highest resolution shell is shown in parenthesis.

CAREERS

BOLD SCIENCE A scientist's move from a devastating loss to a 'brain in a dish' p.283

SCIENCE OUTREACH To engage people, create engaging data go.nature.com/obsud9

NATUREJOBS For the latest career listings and advice www.naturejobs.com



CLAIRE WELSH/NATURE

RESEARCH PROFILES

A tag of one's own

Digital identifiers can sort out different scientists with the same names, and create a lifelong record of their work.

BY QUIRIN SCHIERMEIER

Jee-Hyub Kim knows his way around the scientific literature. The South Korean computer scientist spends his days tracking information and visualizing data drawn from published work in genetics and biomedicine.

Kim works at the European Bioinformatics Institute (EBI) in Hinxton, UK, and knows firsthand what a faceless affair the world's fast-growing scholarly record is. When he searches for his surname in Europe PubMed Central, a chief information resource for biomedical

and health researchers, he gets no fewer than 400,000 hits. If he includes his initials in the search, the system still yields some 15,000 articles — yet only 11 are his.

This ambiguity is hardly surprising. 'Kim' is the most common family name in North and South Korea, and it is a popular given name in other countries. In the global profession of science, similarity between authors' names makes distinguishing researchers difficult for librarians, publishers, funders and administrators. But there is a remedy: the Open Researcher and Contributor Identification (ORCID) project, a

community-driven non-profit collaboration launched in 2012. ORCID provides researchers and scientific contributors with a unique digital identifier that will remain associated with them throughout their lives — even if they change their name or professional affiliation.

Worried about the countless 'false positives' produced by online searches for his research record, Kim did not hesitate to sign up on ORCID last year when he first heard about it. But it wasn't just the ambiguity of his name that prompted him to register: once assigned, the digital identifier allows researchers to manage a record of their activities. Kim's personal ORCID ID thus enables him to link his name not just to papers, but also to other achievements and projects that he has been involved in. "ORCID just makes me and my research profile more visible," he says. "And when I apply for grants or jobs, it is so handy to have all my output filed under one system."

TECHNICAL FINESSE

ORCID is an electronic hub that connects researchers with their research across database profiles, manuscript submissions, grant and patent applications and other such uses. The system's user registry is free, and provides an interface that supports system-to-system communication and authentication.

The platform can, for example, automatically import records from other research-tracking systems, such as Europe PubMed Central and Elsevier's literature database Scopus. This tool allows users to easily collate their publications and make their professional pursuits traceable for potential collaborators, funders, reviewers, employers and colleagues.

Institutions are quickly beginning to see ORCID's value. At the European Bioinformatics Institute, a division of the European Molecular Biology Laboratory, administrators strongly encourage newly hired researchers to sign up for an ID during the induction process. ORCID's ability to facilitate research management and track scientists' output is of considerable value to funders and employers, notes Johanna McEntyre, head of literature services at the EBI.

ORCID is particularly useful to early-career scientists who are seeking to get funded and advance their careers. It makes unpublished yet creditable work more visible. "As modern science gets more and more collaborative, people tend to do a lot of work for which they don't get due credit," McEntyre says. "Results of high-throughput sequencing you have done for genome studies may never get published, ►

► for example. On ORCID, you can claim that data, and draw attention to any other contributions to collaborative research that may help raise your profile.”

To creative minds, assigning a number might conjure up Orwellian associations. But ORCID users have a lot of control. They can choose different levels of privacy for their digital content, and change these settings at any time. They can make some records publicly available and others visible only to trusted parties. And if they wish to list specific works, data or funding sources solely for their own reference, the information can be entirely hidden.

“ORCID is an opt-in system,” says Laurel Haak, the system’s executive director in Bethesda, Maryland. “We do not collect any private information other than e-mail addresses, which the researcher can set as private so that it is not shared.” Privacy concerns have not been a barrier to ORCID’s adoption at the EBI, says McEntyre. “If I really want to find out about you, I just google your name,” she says. “A digital ID doesn’t tell you a lot.”

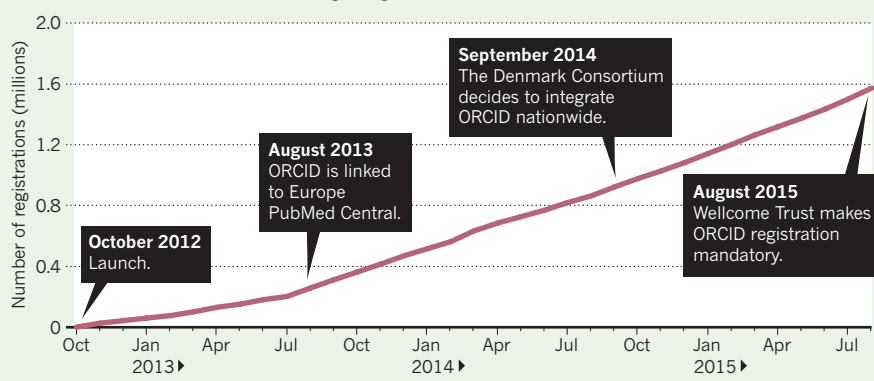
DUAL DIRECTIONS

For the purpose of authentication, ORCID records do display the source — if available — of claimed professional affiliations. If a scientist says that she or he is employed by the University of Oxford, UK, for example, and the university has in turn verified that the individual works for them, that is visible in the system.

With millions of researchers worldwide, and so many similar names, allowing institutions to confirm researcher affiliations helps

NOW 1.5 MILLION STRONG AND GROWING

ORCID’s digital identifiers link scientists with publications and contributions. Sign-ups have increased by more than 50,000 a month since the beginning of 2015.



SOURCE: LAUREL HAAK/ORCID

to ward off concerns that someone could take advantage of name ambiguity. But Josh Brown, ORCID’s regional director for Europe, says that if researchers were tempted to claim papers or data that they didn’t actually produce, it would hardly go unnoticed. “The idea is that ORCID data that is shared between systems is open for validation and cross-checking by those systems,” he says. “By displaying provenance and by the nature of the data itself — which tends to be publicly available or verifiable — any misuse can be detected by the people best placed to do so.”

All 500 or so staff scientists at the EBI have signed up for an ID. So have most scientists at other branches of the European Molecular Biology Laboratory, located in the United Kingdom, Germany, France and Italy. In turn, many science journals (including *Nature*, which partners with ORCID) encourage its use, with the goal of optimizing the manuscript-submission process.

Funding agencies are following suit. To streamline the handling of grant applications, the Wellcome Trust in London has required all applicants since 1 August to provide an ORCID ID. Similarly, the European Research Council, run by the European Commission, has begun asking grant applicants for their IDs — although providing one is not mandatory — so that reviewers can better gauge their skills and contributions to science. And the US National Institutes of Health is testing ORCID’s efficacy for linking researchers and their outputs.

ORCID is rapidly becoming the default global research-management system, says Liz Allen, head of evaluation at the Wellcome Trust. She thinks that scientists should sign up for an ID early in their career and strive to keep their profiles up to date (see “Tips for profile growth”).

“ORCID helps young scientists arrive and settle in the research ecosystem,” says Allen. “It allows you to distinguish your skills from those of co-authors and competitors. And it helps you spend more time doing research and find people to collaborate with, rather than filling out personal information on countless forms.”

Creating a profile is simple. Once a researcher provides a name and an e-mail address, that

individual is assigned a 16-digit number. The ORCID ID is then expressed as a web address to which any publications and personal details can be posted or imported from data repositories.

“It took virtually no time to register,” says Nadarajan Veerapen, a software developer at the University of Stirling, UK, who specializes in process optimization. “Once you’re linked to different systems, ORCID is very practical and easy to handle. But you should really use your ID and not just leave it idle.”

Researchers generally have to maintain an ORCID profile manually. Whether it is acceptable for employers to create ORCID profiles for their staff and feed the system with data from in-house publication databases is still under discussion. “From our point of view, it would make a lot of sense. But there are technical and legal issues that need to be addressed,” says Bernhard Mittermaier, a librarian at the Jülich Research Centre in Germany.

GLOBAL WAVE

ORCID stretches across all disciplines and continents. Worldwide, more than 1.5 million users have signed up since its launch (see “Now 1.5 million strong and growing”).

Not everyone maintains their profiles diligently — in fact, many ORCID users list no output at all. But as the facility gets better known and more widely accepted, the value of its information will increase. As of August, more than 400,000 of the 3.3-million full-text articles listed on Europe PubMed Central were linked to at least one ORCID ID. Some multi-author articles are claimed by several dozen ORCID IDs. And prolific authors and users, such as EBI genome researcher Nick Goldman, have co-written nearly 100 ORCID-claimed papers.

As more people sign up, Allen says, ORCID promises to become a powerful tool for tracking and maximizing the research value of grants, avoiding duplicate funding and identifying opportunities for collaborative research. In June, the British Library in London started to include ORCID IDs in its national thesis service, which provides free access to doctoral theses done at

ORCID BLOOM

Tips for profile growth

- In your profile, list all variations and abbreviations (including initials) of your name, and any previous iterations used in a professional context.
- Carefully choose the privacy settings that best suit your needs.
- Link ORCID with other identifiers and research profiles that you use. For example, if you have a Researcher ID or Scopus Author ID, you can import information from those systems into your ORCID record.
- ORCID can handle a large variety of scholarly output other than scientific articles. Check out what data sets and figures you can link to your profile.
- Make sure to keep your record updated.
- Provide your ORCID ID when submitting manuscripts and applying for grants. Include it in your e-mail signature and CV, and add it to your social-media accounts. **Q.S.**

UK higher-education institutes. In the same month, a group of Italian research organizations announced that it would implement ORCID nationwide, aiming for 80% of Italian researchers to have an ID by 2016.

The system is still developing. To recognize scientists' peer-review activities — time-consuming work that tends to remain invisible — ORCID is discussing with publishers ways to enable scientists to add reviews to their profiles. "We don't get acknowledged" for such work, says Veerapen. "It would be really good if funders and employers were able to check what service I'm doing for science in that respect, too."

There are downsides to using ORCID. The web interface is not perfect, and it is still inconvenient to search, says McEntyre. Others note that ORCID's connectivity with research-tracking systems and databases could be improved. And still missing, Kim says, is a format for registering the software products that often emerge from data-generating research such as his. Haak says that such technical issues will be tackled in consultation with ORCID's 350 or so member organizations, most of which are in Europe, Asia and North America.

But ORCID's strengths — author-name disambiguation and the opportunity to specify unpublished contributions to science — appeal all the same to scientists and research agencies in other parts of the world.

"ORCID is ideal for developing science markets," says Matthew Buys, ORCID's regional director for Africa and the Middle East. "It sits really well with the community in Africa — not just because there are many shared names there, but because funders, publishers and institutions understand that they need to connect to bring high-quality research to Africa."

With 4,500 assigned IDs, South Africa is the best-represented country on its continent. But excitement about ORCID's value is growing in Africa and in the Middle East, Buys says. ORCID's outreach workshops in developing countries — such as one held in July in Nairobi — are well attended.

Ayodele Alonge, a PhD student at the University of Nairobi's School of Journalism and an emerging-technology librarian at the University of Ibadan in Nigeria, signed up for ORCID immediately after learning about it in May. "ORCID enhances my visibility as an upcoming researcher," he says. "And I hope it'll help me get recognized for what I'm doing." ■

Quirin Schiermeier is Nature's Germany correspondent.

TURNING POINT

René Anand

After Hurricane Katrina destroyed his lab at Louisiana State University in 2005, René Anand embraced high-risk research — projects that might win big or fail completely. Anand tells Nature how that decision led him to create what he considers the most-advanced brain model developed so far.

You began a career doing molecular biology.

What sparked your interest in neuroscience?

When I got my PhD at Ohio State University in 1989, I was investigating how genes recombine. I moved on to a postdoc at the Salk Institute for Biological Studies in La Jolla, California, where my informal training in neuroscience started. It was the obvious next frontier in science.

How did Hurricane Katrina affect your lab?

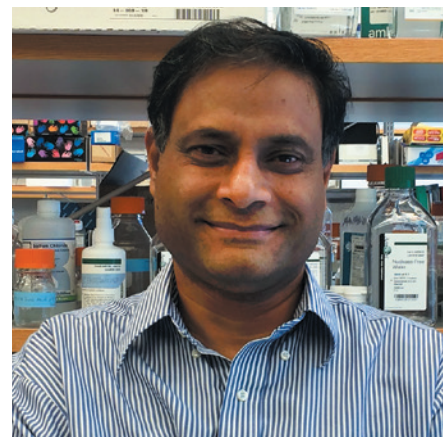
Katrina destroyed the lab itself. The building was flooded with water for a month — the whole system closed down. I lived on the outskirts of town, relying on food from aid organizations and writing grants. A lot of people, mostly clinicians, lost their jobs. Those dark days lasted a year. We chose to move mostly for family reasons — schools were disruptive — not because we didn't foresee recovery. But it was scary to think that something like this could derail us again. Science is already demanding enough. I went back to Ohio State in 2007 — I wanted something familiar and to be part of an interdisciplinary campus.

In 2010, you got a US National Institutes of Health grant. Was that a game-changer?

Yes — I was rewarded for being a risk-taker. The EUREKA (Exceptional, Unconventional Research Enabling Knowledge Acceleration) grant was designed to help investigators to pursue innovative ideas. I wanted to understand at the genomic level how an electric eel's membrane proteins work, so that we could study human diseases involving similar proteins. Getting that grant played a big part in my attempt to turn stem cells into a brain organoid.

How did you decide to create a brain model?

It grew out of my fundraising work with Autism Speaks, a US charity that supports basic research. Year after year, I sat with families and listened to them talk about how much it mattered to them what scientists do. I developed a very personal connection that drove me to take risks. At the time, I was doing research in rodents that failed miserably. I had to find another way, so that I could work in a species that could give us more insight. Using stem cells as the basis for an organoid offered that bridge.



We were fortunate to find two risk-taking funders that gave us roughly US\$140,000. We spent four years producing a stem-cell-based brain organoid using adult human skin cells.

What was the reaction to this 'brain in a dish'?

I should be clear that this 'brain' models early-developmental tissue and is roughly 2–3 millimetres long. It expresses more than 98% of the genes present in a human brain at 5 weeks of development. We are not capable of addressing higher-order function, such as memory, learning or cognition. But we can see structures of the brain, and perhaps use the model to see how it responds to drugs. The organoid might be useful for high-throughput screening for therapeutic-drug discovery or toxicity testing. We are working through legal issues, such as intellectual-property rights. I have the paper ready to submit as soon as we get the business concerns addressed. I didn't realize that the world of commercialization is as challenging to navigate as the science.

How did word about the model get out if the paper is not yet published?

We finished the project in April 2014. As we grew more confident in our results, I shared them at conferences, including an invited talk at the Wellcome Trust in London last July. But it didn't receive press attention until I gave a talk at the Military Health System Research Symposium in Fort Lauderdale, Florida, in August, and my university put out a press release. There are caveats. Although my group has replicated the research, it has not been through peer review. The truth will become the truth once it has been replicated in another lab. ■

INTERVIEW BY VIRGINIA GEWIN

This interview has been edited for length and clarity.

THE MANY MEDIA HYPOTHESIS

How to look after yourself.

BY MARISSA LINGEN

“Are you coming to bed?” he calls.
“In a minute,” you call back. “Someone on the Internet is wrong. And it’s probably me.”

You tell yourself, that is, you #20 — you have a little list by the computer, this is the you with a pit bull — that you really should take the dog in for a vet check with eyes looking like that. Reassured, you scroll down through the rest of the social-media page.

In another world, you had a grandfather who remarried. You’ve seen the pictures of your teenage self on Throwback Thursday, same terrible time growing out your bangs, carting a manic pigtailed toddler around on your hip. In that world you had cousins. Now you’re looking at your cousin’s wedding pictures. She is radiant. She wears flowers in her long blonde hair. You can see where you’re tagged, in the background raising a glass of wine to your other cousin, her brother.

You know that there is a world in which you would give these kids a kidney. In this world, they seem like nice people and you wish them all the best. The echo of that other feeling rings your heart like a goblet.

A handful of you stayed in physics. They are the ones you blame for this. None of your friends has found their social-media pages taken over by alternate timeline selves — although, to be fair, you’ve only trusted a few enough to ask. You still get the posts from your old roommate, your ninth-grade French teacher, all your many cousins in Sioux Falls. They’re just overwhelmed by the sheer number of posts from other worlds featuring, apparently, you.

There is a lot of you.

You have a lot in common.

For example, you never live south of the Iowa border for more than five years as an adult. You don’t all live in the United States or North America or Anglophone countries — but that forty-fourth parallel really seems to be important somehow. You never have more than three kids. You never, ever like coconut, not any of you, not ever.

On the other hand, some of you like pineapple. Some of you even can stand durian, although those are the weird ones and

nobody really understands them. You tend to form clumps. The durian eaters. Those who took to Agatha Christie novels. Those who got good at tennis. Those who let their ear piercings grow shut — that’s not important to the rest of you, but to the ones who did, it’s emblematic of something they can’t explain.

And of course, those who are friends with Stella and those who never met her. Those who lived in Oregon and those who only visited. Those whose hearts were broken when Mrs Bremmer died and those to whom she was just a name on their interdimensional postings.

Those who have developed the full-blown forms of each of three conditions you’re prone to, and those — the majority — who are perfectly healthy. Not you. Those others.

Before the social-media thing you were never very good at taking care of yourself, but you have got much, much better. The latest message you get would have made you weep and rage when all this started. Now you just press your lips together.

You remember breaking up with a guy in college when he passed from annoying to creepy, wondering why you didn’t do it when he got annoying. This other you, you #572, she didn’t even do that well. So she didn’t get the feeling of dodging

a bullet when he pushed her into the wall of the dorm and punched the wall next to her head. She didn’t think, so that’s what my life would have been.

Because that’s what her life is.

You think back to the point where you diverged, to who you know in common with you.

“Go to Heather and Dave,” you write. “They’ll help you get out. They’ve got a spare room, and if they don’t have enough space” — you briefly forget how many kids this you has, if you ever told you — “Heather’s folks are right down the road. They’re good people. They won’t want you to stay in this situation.”

“I haven’t seen them in years!” you write back almost immediately.

Neither have you, but you trust them all the same. Heather was always baffled by your brief relationship with this guy, relieved when it was over. You need a friend like that, and you know you have one — even this many years later.

“It won’t matter,” you write. “She’ll know this is important. Do it for the kids.” You consult your list quickly. Two of them. That’s two too many for this behaviour. “Promise me. Promise yourself.”

The pause is too long. Bad things have happened to you before, but never this. You wonder if you should message another you, but you can’t think which one.

“All right, I promise,” you finally write back. You breathe a sigh of relief. Something might have happened to Heather — it might not work — but you will come back and ask again if you need to. You know you that well.

You are so tired.

You flip back to the wedding pictures, scrolling through. One of your actual aunts is there, wearing a hideous flowered dress. In your own timeline she has developed taste. In the other, not so much. But she has a piece of cake with overblown foliage to match her dress; she has old family friends who in your universe are off doing something else. She looks happy.

That’s all you want for them. All of them. ■

Marissa Lingen has published more than 100 short stories in venues such as Analog, Lightspeed and Tor.com.

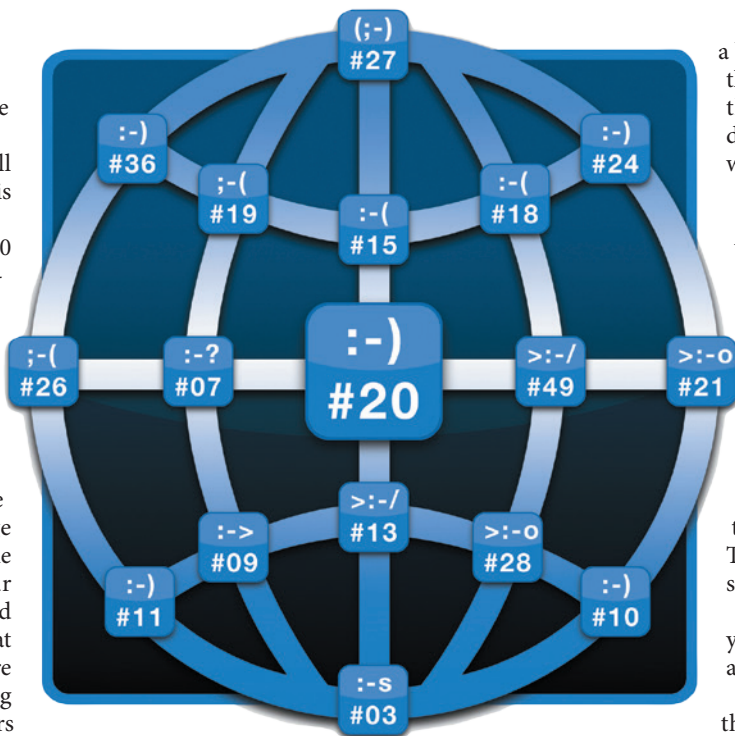


ILLUSTRATION BY JACEY

➔ NATURE.COM

Follow Futures:

🐦 @NatureFutures

📌 go.nature.com/mtoodm

natureOUTLOOK

BEAUTY

8 October 2015 / Vol 526 / Issue No 7572



Cover art: Nik Spencer

Editorial

Herb Brody,
Michelle Grayson,
Jenny Rooke

Art & Design

Wesley Fernandes,
Mohamed Ashour,
Andrea Duffy

Production

Karl Smart, Ian Pope,
Mira Loufti

Sponsorship

Stephen Brown,
Samantha Morley

Marketing

Hannah Phipps

Project Manager

Anastasia Panoutsou

Art Director

Kelly Buckheit Krause

Publisher

Richard Hughes

Chief Magazine Editor

Rosie Mestel

Editor-in-Chief

Philip Campbell

This Outlook is different from most. Instead of focusing on a disease, we move up the hierarchy of human needs above survival, or even health, into the realm of aesthetics. Although beauty could include sunsets and scientific theories, our focus here is on the attraction between humans, and that between other animals that helps to fuel the engine of natural selection.

Neuroscience grants an insight into the traits that have maintained their appeal over the centuries and provides an understanding of how the brain responds to a desirable face (see page S2). In pursuit of beauty, many turn to the products and services peddled by a robust cosmetics industry. A number of these products make scientific claims — some of which are more valid than others (S4). At the more extreme end of the industry, we examine the steady growth of cosmetic surgery. The rising demand for procedures from a more diverse mix of people is leading aesthetic surgeons to rethink facial ideals in a more inclusive way (S6). Men — often neglected participants in the pursuit of beauty — are also starting to get their due (S12). Some people, however, can become obsessed with their appearance, which can lead to a preoccupation with imagined flaws (S14).

Insights into human beauty can be gleaned from researching what it is that other animals find appealing (S8). Evolution has furnished animals with a host of visual cues that signal suitability for perpetuating a species. Stepping back, theoretical physicist David Deutsch makes the case for the concept of 'objective beauty' (S16) and anthropologist Karl Grammer teases apart the role of beauty in human interactions (S11).

We are pleased to acknowledge the financial support of KYTHERA Biopharmaceuticals, Inc., in producing this Outlook. As always, *Nature* retains sole responsibility for all editorial content.

Herb Brody

Supplements editor

CONTENTS

S2 NEUROSCIENCE

The aesthetic brain

Neurological basis for attractiveness

S4 COSMETICS

Molecular beauty

The technology behind skincare

S6 SURGERY

Diverse interventions

The rise of cosmetic procedures

S8 ANIMAL BEHAVIOUR

Come mate with me

Beauty in the animal world

S11 Q&A

Innate attractions

Karl Grammer discusses beauty from an evolutionary perspective

S12 MASCULINITY

Men's makeover

Researchers are forming a clearer picture of male self-image

S14 MENTAL HEALTH

Monsters in the mirror

Research into body dysmorphic disorder

S16 Q&A

Objective beauty

David Deutsch argues the case for objective beauty

S17 BEAUTY

4 big questions

Key research areas

COLLECTION

S18 Rewards of beauty: the opioid system

mediates social motivation in humans
O. Chelnokov et al.

S20 Abnormal brain network organization in body dysmorphic disorder

D. Arienzo et al.

S30 Visual exposure to obesity: Experimental effects on attraction toward overweight men and mate choice in females

E. Robinson & P. Christiansen

S35 Morphological and population genomic evidence that human faces have evolved to signal individual identity

M. J. Sheehan & M. W. Nachman

Nature Outlooks are sponsored supplements that aim to stimulate interest and debate around a subject of interest to the sponsor, while satisfying the editorial values of *Nature* and our readers' expectations. The boundaries of sponsor involvement are clearly delineated in the *Nature Outlook Editorial guidelines* available at go.nature.com/e4dwzw

CITING THE OUTLOOK

Cite as a supplement to *Nature*, for example, *Nature* Vol. XXX, No. XXXX Suppl., Sxx–Sxx (2015).

VISIT THE OUTLOOK ONLINE

The *Nature Outlook Beauty* supplement can be found at <http://www.nature.com/nature/outlook/beauty>. It features all newly commissioned content as well as a selection of relevant previously published material.

All featured articles will be freely available for 6 months.

SUBSCRIPTIONS AND CUSTOMER SERVICES

For UK/Europe: Nature Publishing Group, Subscriptions, Brunel Road, Basingstoke, Hants, RG21 6XS, UK. Tel: +44 (0) 1256 329242. Subscriptions and customer services for Americas – including Canada, Latin America and the Caribbean: Nature Publishing Group, 75 Varick St, 9th floor, New York, NY 10013-1917, USA. Tel: +1 866 363 7860 (US/Canada) or +1 212 726 9223 (outside US/Canada). Japan/China/Korea: Nature Publishing Group — Asia-Pacific, Chiyoda Building 5-6th Floor, 2-37 Ichigaya Tamachi, Shinjuku-ku, Tokyo, 162-0843, Japan. Tel: +81 3 3267 8751.

CUSTOMER SERVICES

Feedback@nature.com
Copyright © 2015 Nature Publishing Group



Portraying Nefertiti as a beauty may have been used as a way to depict the queen's moral qualities.

NEUROSCIENCE

The aesthetic brain

By studying how the brain responds to beauty, researchers hope to understand why we give some people an easier ride or appreciate certain artworks.

BY CHELSEA WALD

In the Neues Museum in Berlin, Queen Nefertiti's head perches, almost weightlessly, on a swan-like neck. The painted stucco and limestone bust is 3,300 years old, but its plump red lips, high cheekbones and almond-shaped eyes look as if they come straight out of a fashion magazine. Indeed, in the 103 years since German archaeologists unearthed the bust, it has achieved an iconic status that supermodels can only dream of. Although a vast chasm of history and culture separates the modern world from ancient Egypt, our continuing admiration of this portrayal lends credence to the idea that some beauty is timeless.

Science has also confirmed the adage, at least to a point. People broadly agree on what faces are attractive, both within and across cultures.

Even babies prefer faces that adults judge to be attractive, suggesting that there is something hard-wired about these preferences. Our judgement of other people's attractiveness often happens subconsciously and influences us in ways we do not realize. Psychologists have observed that citizens vote for more attractive political candidates, judges give attractive defendants more lenient sentences and teachers grade better-looking students more favourably (see 'Snap judgement'). "The breadth of circumstances that seem to be affected by facial attractiveness is mind boggling," says psychologist Benedict Jones of the University of Glasgow, UK.

These observations have been difficult to explain, Jones says. But now, using technologies that range from digital face morphing to brain imaging, psychologists and neuroscientists are starting to identify the diverse qualities that humans find attractive in faces, as well as the

complex networks in the brain that respond to beautiful features. Their work is not only uncovering neural links between evaluations of attractiveness and those of social attributes such as trustworthiness, but is also giving an insight into our appreciation of artworks such as the Nefertiti bust. "This isn't a trivial quirk of our facial structure," says neuroscientist Peter Mende-Siedlecki at New York University. Beauty may be difficult to define, he says, but it is real and its influence is vast.

WHY BEAUTY?

Thinkers and artists throughout history believed that facial beauty was intrinsically linked to certain ideal proportions — to Plato, for example, the width of a face should be two-thirds of its height. In fact, it is much more complicated than that. Starting in the 1990s, psychologists began to adapt special-effects techniques such as digital morphing to construct faces that people found more attractive. They identified three key qualities — symmetry, sexual dimorphism (femininity and masculinity) and 'averageness' — that correlate with attractiveness, says Jones.

Why these traits? In the case of symmetry and highly feminine traits, scientists posit that we may have evolved a preference for them. "These are things that are thought to be quite important for mate choice and mate preference in many non-human animals," Jones says. A symmetrical face may indicate a healthy development, free of genetic disorders or infectious diseases. A feminine face — think of Nefertiti's lips, cheekbones and eyes — could indicate fertility. And in fact, data suggest that feminine features are linked with higher oestrogen levels and hence with fertility. For masculine traits, however, psychologists speculate that a hyper-masculine face may be a 'costly signal' — a sign that a man has energy to spare. Many studies have shown that women prefer more masculine characteristics around the time of ovulation.

The story behind averageness is less straightforward. In the late 1800s, Victorian polymath Francis Galton invented a way to make composite portraits by superimposing photographs of different people. He hoped that this would help to identify the common physical characteristics of criminals, of those with a disease or of other 'biological types' such as a person's ethnicity. But in the process he noticed that the composites were generally better looking than the individuals who made them up. "The special villainous irregularities" had been removed, he wrote. A century later, psychologists followed up on Galton's observation¹, using digital techniques to show that people do indeed find these averaged faces to be more attractive than the originals.

Averageness, like symmetry, may be a signal of health — in particular, a lack of potentially harmful genetic irregularities. But this aspect of beauty might instead be a by-product of how the brain works, says Marcos Nadal, a psychologist at the University of Vienna who

studies aesthetic experiences. Processing faces is an exceedingly complex task. The brain might just prefer faces that resemble the average face of a population because they are easier to identify than less typical faces. “There are studies that show that averageness plays a role in the attractiveness of many other objects,” like an average-looking watch face for example, says Nadal. “The brain works by extracting regularities.”

For now, these explanations for our preferences are conjectures. And as scientists gather further evidence, the picture of why we find others attractive grows more complicated. In one study, for example, researchers found that some non-industrialized societies do not consider highly feminine or masculine faces to be especially attractive². The authors speculate that a preference for sexual dimorphism could arise from processing lots of diverse faces in a densely populated environment. Skin quality, fat distribution and expression can also contribute to attractiveness, supporting the idea that beauty is linked to many markers of fitness, as well as a potential mate’s receptiveness. “Even making what may seem like a simple judgement — is this face attractive or not? — is dependent on a very complex system involving many different inputs,” Mende-Siedlecki says.

THE ROOT OF ATTRACTION

This beauty-recognition system is part of the neural network that processes faces, which is shared between various brain regions. The occipital lobe, at the back of the brain, receives the signals from the eyes. Here, specialized areas extract basic information about the face being observed, such as features, expression, eye gaze and lip movement. These data are then bounced forward to the parts of the system that process higher-level information such as emotional state.

Studies consistently show that attractive faces light up the brain’s dopamine-driven reward network. For example, researchers have found people would press a key to see an attractive face for longer, in much the same way as a mouse will press a lever to get food or drugs, Jones says. Those faces stimulate areas such as the nucleus accumbens, which Nadal calls “a generator of pleasurable sensations”.

A key module of this system is the orbitofrontal cortex, which sits just above the eyeballs. Neuroscientists think that the middle part of this region is where the brain judges the value of a potential reward. In a study in which people could win money, this area showed more activity when the winnings were bigger³. “It’s focused on attractiveness with a positive bent: this is something of great social value,” says Mende-Siedlecki. Researchers have also shown that a different part of the orbitofrontal cortex — an area that is associated with punishment — responds to unattractive faces. So just as seeing an attractive face may feel like winning money, seeing an

SNAP JUDGEMENT

We infer social traits just by looking at a person’s face — attractive features are associated with being trustworthy.



unattractive face may feel a little like losing it.

Attractiveness activates these reward areas even if we are not consciously thinking about the beauty of a face. Outside of the reward network, this is also true for some core parts of the face-processing system in the visual cortex. Neurologist Anjan Chatterjee of the University of Pennsylvania in Philadelphia and his colleagues showed participants 100 different images of faces, asking them to evaluate either the faces’ attractiveness or their identities. Using functional magnetic resonance imaging, the researchers found that brain areas specializing in face recognition showed more activity when participants looked at faces that they had previously rated as attractive than when they looked at less appealing faces⁴. Such enhanced activity occurred even if participants were thinking only about the face’s identity and not about its attractiveness. This shows that the brain responds rapidly and automatically to beauty, Chatterjee says — even when beauty is not on our mind.

FEAR FACTOR

Research on facial attractiveness is also leading neuroscientists to re-evaluate an almond-shaped emotion centre deep in the brain, called the amygdala. “For so long, we thought the amygdala is all about threat, all about snakes and spiders,” says Mende-Siedlecki. Indeed, early studies focused on the role of the amygdala in processing fearful faces. But it is now clear that the amygdala reacts to all kinds of faces. A few studies have also indicated that, unlike other regions such as the orbitofrontal cortex, the amygdala may respond to attractiveness in a non-linear way — the reaction gets stronger the more beautiful or ugly a face is, and weaker for more neutral-looking visages. “The signal is saying, there’s something here that’s kind of weird, kind of unexpected, not what I’m used to,” Mende-Siedlecki says.

The amygdala also contributes to judgements of trustworthiness, says Mende-Siedlecki. This overlap might be efficient for the brain, but as a side effect it could play a part in what psychologists call the attractiveness halo effect — a reflexive presumption that external beauty indicates overall goodness. Such a neural short cut can lead to all sorts of social benefits for attractive people, from better

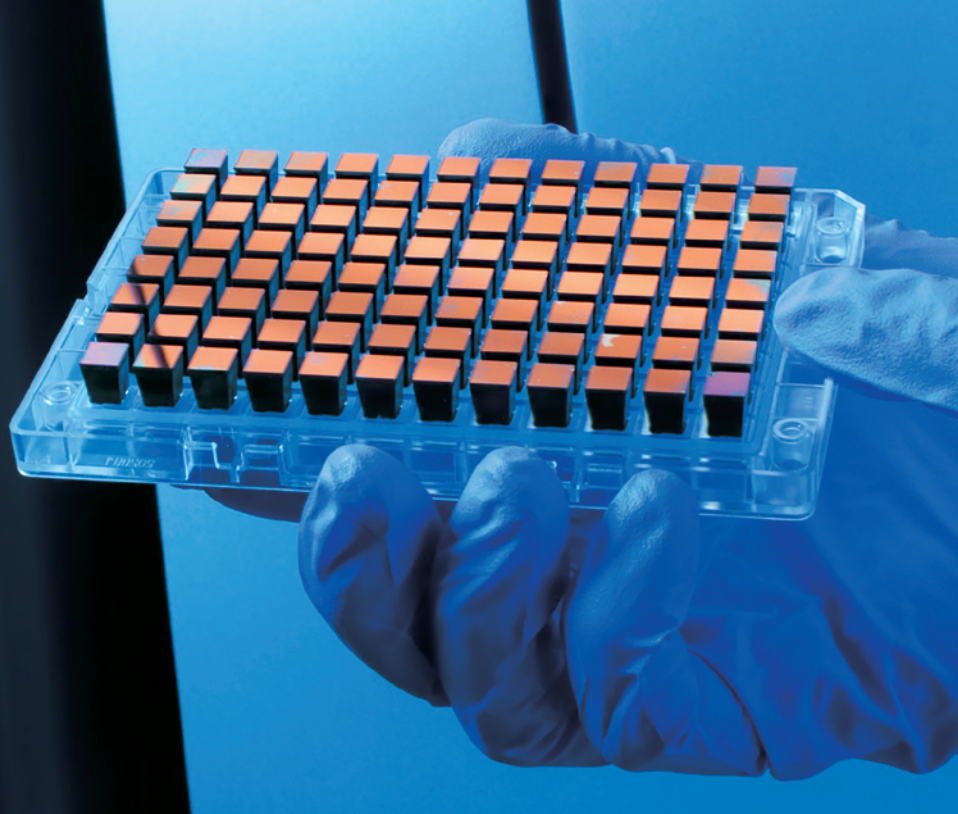
grades to more lenient punishments. Jones and his team have found that people presented with hiring scenarios are more likely to award higher salaries to more attractive people⁵. He contends that society should address this bias as much as it does for other prejudices. “Hopefully, you will start to see things like people taking into consideration how you can minimize the effects of facial appearance on court decisions and hiring decisions,” he says.

The study of facial attractiveness is also helping neuroscientists to start to understand a completely different aspect of society: aesthetics. “In so far as we understand something about the neural response to beauty, we can begin to generate hypotheses about neural responses to other objects, including art objects,” says Chatterjee, who is also the author of *The Aesthetic Brain* (Oxford Univ. Press, 2013). Take Nefertiti’s bust, for example. Academics agree that it probably does not resemble the real queen. Among other things, the sculptor Thutmose rendered it remarkably symmetrical.

Why would he do that? “People have used the metaphor of artists as intuitive neuroscientists, in the sense that they have been able to engage the brain mechanisms that make people become interested or shocked or enamoured,” says Nadal. He speculates that Thutmose was not after realism or even physical attractiveness when he made his masterpiece. Instead, he was taking advantage of the halo effect, accessing the deep link that the brain makes between beauty and other virtues. “Beauty would convey Nefertiti’s moral qualities, like goodness or justice or rectitude,” Nadal says. And as we know from historical accounts of her peaceful and prosperous reign, this queen was much more than just a pretty face. ■

Chelsea Wald is a freelance science writer in Vienna, Austria.

- Langlois, J. H. & Roggman, L. A. *Psychol. Sci.* **1**, 115–121 (1990).
- Scott, I. M. et al. *Proc. Natl Acad. Sci. USA* **111**, 14388–14393 (2014).
- O’Doherty, J., Kringelbach, M. L., Rolls, E. T. Hornak, J. & Andrews, C. *Nature Neurosci.* **4**, 95–102 (2001).
- Chatterjee, A., Thomas, A., Smith, S. E. & Aguirre, G. K. *Neuropsychology* **23**, 135–143 (2009).
- Fruhen, L. S., Watkins, C. D. & Jones, B. C. *Leadership Quart.* <http://doi.org/7q3> (2015).



Gene-expression analysis is allowing researchers to define the pathways associated with ageing.

COSMETICS

Molecular beauty

The rise of genomic and other technologies in cosmetic skincare is leading to products that might improve skin health.

BY ALLA KATSNELSON

As scientists raced towards the finish line of the Human Genome Project at the turn of the twentieth century, a New York-based university spin-off called Lab21 set out to apply genetic-sequencing technology to skincare. Using a 'skin DNA test' that assessed mutations in five genes, the company claimed to have designed personalized skincare concoctions that would moisturize, plump and de-wrinkle any individual's face — at a cost of US\$250 for a month's supply. "We have taken the guesswork out of the skincare equation," the company's president proclaimed in a 2003 press release. Upscale department stores eagerly signed on.

Geneticists and dermatologists scoffed at Lab21's researchers, saying that too little was known about the genes involved and about how a cream's active ingredients might supplement genetic failings. They turned out to be right. Lab21 fizzled out after a few years, but it was not the last cosmetics company to grab on to the coat-tails of scientific advancements.

Claims of scientific efficacy are so common for today's skincare products that they elicit eye-rolls from the sceptical. But such claims were not always the norm: until a couple of

decades ago, "the beauty industry was almost allergic to science", says Barbara Gilchrest, a dermatologist at Massachusetts General Hospital in Boston. The common perception, she says, was that consumers were scared of and uninterested in science, so companies did not want the word associated with their products. That aversion started to melt with the advent of genomics, which not only revolutionized research but — as the enthusiasm for Lab21's product suggested — also captured the public's imagination. Of course, outsized interest has fuelled some "rather flagrant pseudoscience", Gilchrest points out. But alongside the flood of dicey assertions, the field is finally getting onto firmer scientific ground, she says, laying the foundation for formulas that may have the power to legitimately reverse skin ageing.

Over the past decade, cosmetics companies have invested heavily in molecular and genomic research into what causes skin cells to age, with the hope of pinpointing ways to interfere with that process. Researchers are applying these tools from the other direction to determine whether already available treatments that seem to work cosmetically also improve

the functional qualities of the skin. "Over the years, we've converged on an understanding that we should be doing really deep biology on the skin-ageing process and on products that can be used to improve skin health," says molecular biologist Jay Tiesman, who works on the beauty brand Olay at consumer-goods company Procter & Gamble. "It's just like any other biological endpoint that a pharmaceutical company would go after."

PROVEN REJUVENATORS

As skin ages and is exposed to ultraviolet light, collagen — the key protein in maintaining skin's elasticity and structural integrity — begins to fragment. Meanwhile, skin cells called dermal fibroblasts that normally produce collagen become less efficient at doing so. Wrinkles, sagging and uneven pigmentation are the result. The first substance demonstrated to treat wrinkles was a vitamin A derivative cream called tretinoin, co-invented by dermatologist Albert Kligman, who was also the first to show that ultraviolet light causes wrinkles. Marketed as Retin-A, the cream was approved to treat acne in 1971, but soon gained a reputation as a wrinkle-buster, and physicians began prescribing it off-label. A small, but influential clinical trial in 1988 demonstrated its efficacy, whipping consumers into a tretinoin frenzy¹. "It really blew open the translational research field for the reversal of skin ageing," says Sewon Kang, a dermatologist at Johns Hopkins University in Baltimore, Maryland. "Up until then, most physicians thought that if the skin starts to sag, you go and find a good plastic surgeon."

Since then, researchers have found that tretinoin cream (and its related compounds) stimulates fibroblasts to make procollagen (collagen's precursor) and supports the skin's extracellular matrix, countering some of the destructive effects of ultraviolet light. But how exactly this happens — and whether it reverses the degradation that occurs with skin ageing — is unknown. Anne Lynn S. Chang, a dermatologist at Stanford University School of Medicine in California, recently embarked on a project to examine how commonly used topical skin products such as tretinoin might change gene-expression signatures and other molecular markers, and to determine whether those treatments have real benefits for skin health, not just appearance. Chang hopes the project will be as productive as a pilot study she conducted, which examined the efficacy of another widely used dermatological procedure called broadband light (BBL) treatment².

To administer BBL, a clinician passes a wand that pulses high-intensity visible and infrared light over a person's skin in a series of sessions weeks apart. The technique has been approved by the US Food and Drug Administration to treat skin discolouration; dermatologists also use it off-label for skin rejuvenation. Chang and her colleagues found that when they used

➔ **NATURE.COM**

Read more about lasers and dermatology at:
go.nature.com/bmbjgq

BSIP SA/ALAMY

BBL on people who had substantial sun damage more than half of the genes whose expression had been altered by age were restored to expression levels similar to those of skin from younger individuals.

In a similar vein, Frank Wang at the University of Michigan in Ann Arbor and his colleagues have explored the efficacy of injecting 'dermal fillers' — specifically, those containing hyaluronic acid. This naturally occurring substance is a key component of the extracellular matrix. Skin researchers have surmised that injecting such fillers tightens sagging skin and smooths wrinkles simply by physically adding volume to the skin, but Wang's group found that the effects go much deeper. Hyaluronic acid injections boost gene and protein expression of type I collagen (the most abundant collagen in human skin) within four weeks³. In further explorations, they reported that this filler — by providing structural support to the extracellular matrix — activates dermal fibroblast cells and stretches them out. This stretching switches on a signalling pathway that stimulates the skin to rev up its own collagen production. Because newly formed dermal collagen persists for many years, the treatment provides long-lasting effects⁴. "That's telling us that fibroblasts don't inherently lose their function with age," says Wang.

A PERSONAL TOUCH

These lines of study point to biological ageing processes that are, at least to some extent, reversible. Pinning down the molecular signalling that drives these processes should unveil approaches for designing new products, says Chang. Large-scale gene studies are beginning to tease out the key pathways that are involved in skin ageing — an approach that is old hat in the pharmaceutical industry, but that has more recently infiltrated the skin-health and cosmetics world. In a study of 428 centenarians, 6 gene mutations were found to correlate with youthful skin appearance. But the mutated genes, it turned out, were not the same ones that are associated with longevity⁵. "One of the significant gene variants is near a gene that is found in immune cells in the top layer of skin," Chang says. This observation suggests that individuals with younger appearing skin may have a different immune response than do other people.

Chang is also collaborating with skincare company Nu Skin based in Provo, Utah, to identify gene-expression signatures in the skin of women who naturally have more or less youthful-looking skin. The study, presented at the Annual Meeting of the Society for Investigative Dermatology in May, identified several hundred genes that differed — many of which are involved in known ageing pathways⁶.

Other companies are hot on the same trail. Alexandra Kimball, a dermatologist at Massachusetts General Hospital, is collaborating



Broadband light therapy involves passing high-intensity light over the skin to treat skin discolouration.

with Procter & Gamble and is thick in the middle of what Tiesman calls "the Manhattan Project of skin ageing". Kimball and her colleagues examined gene-expression patterns in 3,700 samples from a total of 225 African American and white women in different decades of their lives, as well as microbial composition, proteomes and metabolomes in a smaller subset; they looked at skin from exposed parts of the body as well as from the usually shaded buttock area.

At the World Congress of Dermatology in June, the researchers reported distinct patterns of gene activity that are characteristic of each decade, as well as the expression pattern signatures of people whose skin aged especially well or especially poorly. One insight the study revealed is that timing and patterns of molecular changes correspond to known broadly age-related alterations — cell senescence begins to appear in the 40s, and the skin's ability to maintain moisture levels starts to wane in the 50s. "With these data," Kimball says, "we can certainly anticipate skincare products personalized for people by decade."

And, coming full circle to the Lab21's offerings of a decade ago, a handful of boutique skincare manufacturers are again offering personalized creams based on DNA testing. One of these is GeneU, founded by Imperial College London engineer Christofer Toumazou. In its London store, GeneU offers microarray tests that assess three variants in each of two genes: *MMPI*, with variants indicating whether a person is a fast, medium or slow degrader of collagen; and *NQO1*, with variants pointing to cells' capacity to fight oxidative stress. Based on the results and on a lifestyle survey of factors such as sun exposure, smoking habits and stress levels, customers receive one of 18 formulations of

the cream. (The initial test plus a 2-week supply costs £600, US\$930; a subsequent supply is £250 per month.) Toumazou says that users experienced a 24–29% reduction in different types of wrinkles, as assessed by dermatologists in a 12-week placebo-controlled, double-blind study.

Those results have not been published, however, and many researchers do not buy the company's claims. "These endpoints have a lot of 'kitchen logic' to them; collagen breakdown is an important part of skin ageing," says Tiesman. But no specific variants in the genes have been explicitly linked to skin health, so it is unclear what actionable information the tests provide.

There is no telling whether the current wave of research will yield cosmetic skincare products that are proven to truly halt or even reverse the ageing clock. Differences in gene expression and other markers are just a start; they must be followed up with studies that explore the underlying biology of the skin, assays of active ingredients and numerous other steps. Just as in drug discovery, dead ends will abound. But these first stabs at building a comprehensive picture of skin ageing have already brought the cosmetics field to a point at which products backed by science, as much as by marketing, seem like a realistic possibility. "It better pay out in the end," says Tiesman, "because it has cost us a lot to execute." ■

Alla Katsnelson is a freelance science writer in Northampton, Massachusetts.

1. Weiss, J. S. et al. *J. Am. Med. Assoc.* **259**, 527–532 (1988).
2. Chang, A. L. et al. *J. Invest. Dermatol.* **133**, 394–402 (2013).
3. Wang, F. et al. *Arch. Dermatol.* **143**, 155–163 (2007).
4. Quan, T. et al. *J. Invest. Dermatol.* **133**, 658–667 (2013).
5. Chang, A. L. et al. *J. Invest. Dermatol.* **134**, 651–657 (2014).
6. Xu, J. et al. *J. Invest. Dermatol.* **135**, S28–S48 (2015).

황금비율에 맞는 디자인으로
최상의 아름다움을 찾아 드립니다.

半永久化妆, 纹唇, 眼线, 可维持5-7年

연예인, 방송인이
즐거 찾는 **메이크업**

생얼로 당당한 이유?

퍼머넌트 메이크업
02-3445-5597
예약문의 010-4490-8020
3호선 신사역 4번출구 3분거리
NAVER 미자민 메이크업 검색

The number of people undergoing cosmetic procedures, such as those advertised in South Korea, is rising among all ethnic groups.

SURGERY

Diverse interventions

Standards for cosmetic surgery are typically based on white ideals of beauty. But the demand for facial procedures by people of all ethnicities is driving a change in practices.

BY SUJATA GUPTA

Around 1,000 years ago, Leonardo da Vinci divided the face into horizontal thirds and noted that the distance from the hairline to the brow, the brow to the nostril, and the nostril to the chin should all be equal. It was one decided moment in a long, obsessive search to find objective ways to classify beauty.

As a young medical student many centuries later, Jennifer Parker Porter recalls sitting in a plastic-surgery class being asked to analyse faces and noses using such historic ideals. “I started thinking, ‘Well, this isn’t right. You’re analysing my nose as a Caucasian nose and I’m not Caucasian,’” says the facial plastic surgeon from Chevy Chase, Maryland.

And so Porter began to take measurements directly from the faces of people of varying backgrounds to quantify racial differences. Her work has put her at the leading edge of a larger effort within the plastic-surgery field over the past few decades — to consider the concept of beauty in less relentlessly Western terms. “The aesthetic ideal, it comes from many moons ago when the

anatomists and artists of yesteryear were looking at proportions of the face and deciding what was normal,” she says. “But they were looking only at Caucasian faces.”

That is now changing. However one feels about cosmetic enhancement through surgical means, the fact is that more and more people of all ethnic groups are having cosmetic surgery. Between 2005 and 2014 in the United States, cosmetic procedures — ones done for aesthetic enhancement rather than for reconstruction, birth defects or diseases — jumped by 38% in white people and 110% in non-white people (chiefly Hispanic, African American and Asian American people), according to the American Society of Plastic Surgeons (see ‘Rise of surgery’).

As this cosmetic-surgery landscape shifts, the reigning aesthetic ideal that surgeons work to has come under intense scrutiny. Reshaping facial features for people of different ethnic groups means re-evaluating the prevailing aesthetic standards and establishing new, more diverse surgical guidelines.

Operating on an individual without considering their ethnic background, says Julius Few, a facial plastic surgeon and clinical professor at the University of Chicago in Illinois,

“is like trying to do heart surgery without knowing where the blood vessels go”.

END OF THE MASQUERADE

Pride in ethnic identity has helped to spur this change. Until the 1990s, the small number of Asian and black people who pursued cosmetic surgery mostly did so to efface rather than celebrate their features. “There was a time when most African Americans were really trying to achieve a much narrower nose, a Caucasian-like nose,” says J. Regan Thomas, a plastic surgeon in Chicago.

Many surgeons report that Asian and black people, for example, now want to preserve their ethnic features. David Weeks, a facial plastic surgeon in Atlanta, Georgia, recalls a 28-year-old Middle Eastern woman he saw a few years ago. She was considering having a rhinoplasty, and Weeks assumed that she wanted to remove the generous hump on her nose, which is a common procedure among his white patients. But she just wanted work on her nose tip. When Weeks delicately asked about addressing the hump, her mother — who had come to the consultation — cut the conversation short.

Other surgeons report similar discussions. “Some Middle Eastern patients might come in

with a huge hump. They might want to soften it, make it a little less out of control, but they feel if they go too far it makes them look Caucasian,” says David Kim, a facial plastic surgeon in San Francisco, California.

Sorting out what it means to be a beautiful Middle Eastern, Asian or black person in contemporary communities means crafting templates of beauty that strip away historical white biases. To do that, Porter and other researchers are trying to quantify what makes different kinds of faces beautiful.

Chung H. Kau, an orthodontic specialist at the University of Alabama at Birmingham, has been doing so by constructing 3D faces — a procedure that entails measuring facial attributes such as the distance between the eyes or differences in the curvature of the nose¹. So far, he has collected data sets from 15 countries to create an ‘average face’ for each. “We do see certain characteristics in certain populations,” he says. He hopes that surgeons will one day design a nose or eye using the appropriate geographical template.

Kau’s work speaks to another thread of research, showing that when several faces are melded into a single composite face, people across the world find the composite more attractive than any of the original faces². Paradoxically, therefore, average equals beautiful. Averageness can be more important in certain features than in others³, however, and retaining a striking feature could actually accentuate beauty.

Few has also delved into how to quantify ethnic differences. After he started practising medicine 16 years ago, among other procedures he looked at surgery that is done to lift the corners of the eyes (which tend to droop with age) to make a person look younger. Several years ago, he reviewed the photographs in his archives and selected 296 white and African American patients, divided into over-45 and under-45 age groups, and compared how the eye aged over time⁴.

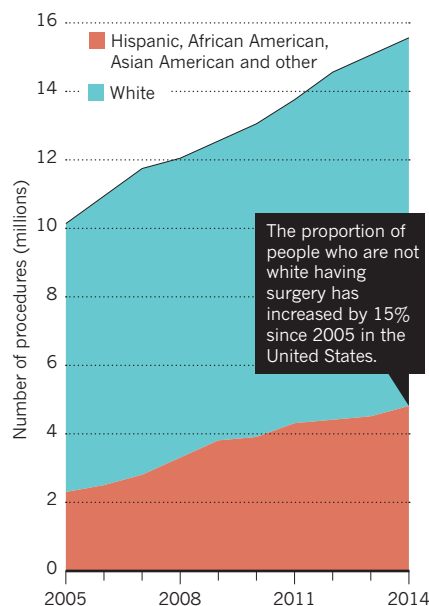
He saw that the African American eye has a greater natural slant, even in the younger age group. Under-correcting the slant in African Americans, he realized, may make the face appear not younger — but whiter. That is important, he says: a mismatch between the surgery and face type can leave patients dissatisfied or confused by their new appearance, even if they cannot articulate what feels wrong. “Even though it’s one small thing, the implications are huge,” he says.

PARSING ETHNICITY

Still, attempts to characterize beauty by skin colour alone come with pitfalls, says Ronald Eccles, a rhinology researcher at Cardiff University, UK. Plastic surgeons often use skin colour as a proxy for race, he notes, but “what is African American?” Or Asian, white or Hispanic person? “You can’t define such terms through

RISE OF SURGERY

Although white people make up the largest proportion of US patients undergoing cosmetic procedures, surgery is increasingly popular among Hispanic, African American and Asian American people.



looks alone.” During the cosmetic surgery consultation, a surgeon must consider facial architecture, such as the height of the nose bridge, curvature of the jaw and skin type (white skin is typically thinner than black skin) to determine the types of surgical fixes that are possible. Beyond those physical limitations, cosmetic surgery is interpretive: how much should the nose be narrowed or eye lifted, for example?

The answers may rest on a person’s ethnic background, which is not always clear from their skin colour. If, for example, a patient appears to be or identifies as Asian, but is three-quarters white or another permutation, a surgical assessment could become skewed: the surgeon might suggest a narrower, more ‘white’ nose when merely removing a hump might work best. Similarly, a surgeon working with a patient who identifies as black may suggest an ‘African American eye’ even though their face type would support a more characteristically white eye.

To avoid ambiguity, some surgeons routinely ask patients about their heritage. For noses, Eccles suggests sorting people not by their skin colour, but by a quantitative assessment of nose shape. This tool was first suggested more than 100 years ago by French physician and anthropologist Paul Topinard, who divided the width of the nose from its widest portion at the base by its height and called the result the nasal index. Although Topinard used the index to make judgements about race, removing this aspect could make the index an objective tool, Eccles says.

Nonetheless, many surgeons say that they do not bring up a patient’s heritage and merely hope that such information will come up during the consultation. But cosmetic surgeons are also

artists whose own life histories will undoubtedly shape their conceptualization of beauty. “I’m probably influenced by my background,” Weeks, who is white, acknowledges. “What I think of as normal may be different from what other surgeons think of as normal”. When those subtle assumptions play out on the operating table, they can dramatically influence the look of a face. Without careful consideration of a patient’s desires and background, some people will be so unhappy with the result that they will resort to further surgery.

TOWARDS BROADER STANDARDS

Beyond questions of aesthetics, cosmetic work is still difficult surgery. Tweaking a nose or chiselling down the jaw — a popular procedure in South Korea — takes considerable finesse. And because cosmetic plastic surgery was once the almost exclusive domain of white patients and surgeons, European and North American doctors have largely laid out the surgical guidelines for various procedures.

That Western orientation is problematic, says Yong Ju Jang, a rhinoplasty specialist at Asan Medical Center in Seoul. Because the writers of the guidelines tended to overlook the nuances of operating on anyone who is not white, different — and often contradictory — standards for surgeries have arisen across the world. In practice, that means a Korean American woman pursuing a nose job may well undergo a different procedure with different aesthetic outcomes in the West than she would in the East.

Consider the nose, says Jang. In general, most white patients wish to reduce its size, whereas Asian patients seek to enhance it to make it appear ‘stronger’. For rhinoplasties in most of his Korean patients, Jang says he builds up the nose using cartilage from the patient’s ribs or ears, as well as using implants made of Gore-Tex or silicone to improve the aesthetic end result. Although Gore-Tex implants are becoming more common for rhinoplasties across Asia, most Western surgeons avoid them for fear of complications, such as infections. Jang, who has been educating surgeons around the world on how to conduct rhinoplasties in Asian patients, says that with practice, those risks are largely eliminated.

Cosmetic surgeons hope that as long as more people of different backgrounds seek aesthetic plastic surgery, the standards will continue to evolve to reflect every type of nose, eye and chin. After all, says Parker, “you can’t just run in and do the same nose on everybody.” ■

Sujata Gupta is a freelance science writer based in Burlington, Vermont.

1. Kau, C. H. et al. *Am. J. Orthod. Dentofacial Orthop.* **137**, S56.e1–S56.e9 (2010).
2. Baudouin, J. Y. & Tiberghien, G. **117**, 313–332 (2004).
3. Weeks, D. M. & Thomas, J. R. *Facial Plast. Surg. Clin. N. Am.* **22**, 337–341 (2014).
4. Odunze, M., Rosenberg, D. S. & Few, J. W. *Plast. Reconstr. Surg.* **121**, 1002–1008 (2008).



The pattern and colouring of male golden-eyed reed frogs (*Hyperolius ocellatus*) may help them to entice females.

ANIMAL BEHAVIOUR

Come mate with me

In a cut-throat world where only the fittest survive, beauty seems to be a needless expense. But creatures are strutting their stuff in ways that help to perpetuate their species.

BY AMY MAXMEN

On São Tomé and Príncipe, two tiny islands off the west coast of Africa, Rayna Bell came across Lilliputian frogs, decked out in lime green, speckled in leopard print and daubed with indigos. South and Central America is home to similarly colourful tree frogs, but the golden-eyed reed frogs that Bell saw are not closely related to these — the brilliant patterns had evolved independently multiple times. To Bell, an evolutionary biologist at Cornell University in Ithaca, New York, that fact suggests that such beauty cannot be accidental. “You can’t look at these frogs, and not think that something is going on,” she says. The question of what that something is drives her research today.

More than 150 years earlier, Charles Darwin had been similarly perplexed by beauty. The trait takes energy to produce and it makes prey easier to spot. A fluorescent orange tree frog

stands no chance of blending into the jungle’s backdrop. In a letter to his colleague, the botanist Asa Gray, Darwin wrote, “The sight of a feather in a peacock’s tail, whenever I gaze at it, makes me sick!”. After years of observations, Darwin proposed an evolutionary concept to account for beauty: sexual selection. Whereas natural selection allows only those who survive to adulthood to pass their genes on to their offspring, sexual selection permits certain individuals to find mates more often than others.

However, strong evidence for Darwin’s sexual-selection theory, and an understanding of how it functions, has emerged only in the past decade. This is because many characteristics are nearly impossible for humans to see, and are discovered only by dissecting the sensory systems of the potential mates of every stylish species.

NATURE.COM
Read more about sexual selection at:
go.nature.com/mmmycbr

demonstrated how such visual preferences can lead to entirely new species, and how beauty is meaningful — not just skin deep.

FLASHING THE OPPOSITE SEX

In Darwin’s 1871 book proposing sexual selection, *The Descent of Man, and Selection in Relation to Sex*, the words ‘beauty’ and ‘beautiful’ appear 280 times. Males, Darwin noticed, tended to be the more flamboyant sex in the animal kingdom, and he supposed females were the pickier. However, he did not have strong evidence to demonstrate that females chose aesthetically pleasing males over duller suitors, and he was not sure what benefits decorated males might confer. Acknowledging that he could not fully justify aesthetics, he flirted with the idea that beauty is a by-product. “Hardly any colour is finer than that of arterial blood; but there is no reason to suppose that the colour of the blood is in itself any advantage; and though it adds to the beauty of the maiden’s cheek, no one will pretend that it has

been acquired for this purpose,” he wrote¹.

But Darwin’s gaze was limited. Although he travelled, and consulted with zoologists and botanists around the world, no one really knew how varied the senses of sight, smell, touch, sound and taste were among creatures. For example, photosensitive cells in human retinas, called cones, see daylight wavelengths ranging from violet to red, whereas the eyes of some insects see deep into the ultraviolet. And although insect vision is similar to that of humans, inasmuch as they see the world through a series of snapshots that the brain weaves together, insects have a faster rate of capture. As a result, insects perceive much smaller fluctuations in movement than we do. And it turns out that courting insect males use this ability to their advantage.

Biologists have long supposed that the decorated wings of male butterflies attract mates. Indeed, one experiment conducted in the 1950s showed that female *Hypolimnys misippus* butterflies preferred males with round, baby blue spots on their hind wings to males with those spots artificially blacked out. Realizing that butterflies rarely sit still, Darrell Kemp, an evolutionary biologist at Macquarie University in Sydney, Australia, wondered whether there was more to the story.

Kemp and his team analysed videos of males and females of a related species, *Hypolimnys bolina*, kept in large cages. They observed how males fluttered below the females that they courted, so that when light reflected off the spots on the males’ wings, it struck female eyes at an angle that made the spots luminesce with ultraviolet light. In addition, the courting males beat their wings at a shallower amplitude and faster than they did while they were foraging. This had the visual effect of transforming the spots into quick bursts of light when viewed from above, flashing about 11 times per second². “If you want to really impress a female — or rather, impress her visual system — the best way is to present a bright colour that flashes on and off,” says Kemp.

ALTERED SUNBEAMS

Fish are similarly susceptible to bedazzlement. A mirror-like layer in their eyes bounces light back through the retina, in such a way that photons have a second chance to be captured by photosensitive cells. As a result, a bright flash (as opposed to, say, a steady beam) draws a fish’s attention. This is particularly true when the flash contrasts sharply against a murky underwater background.

Molly Cummings, an evolutionary biologist at the University of Texas at Austin, suspected that northern swordtails, *Xiphophorus nigrensis* — which bounce light of their shiny silver skin in the same way that glaring polarized sunlight is reflected off a lake — might harness polarized light to attract the opposite sex. To test that idea she and her colleagues filmed the

male swordtails in tanks as they swam beside females. By altering the type of artificial light in the tanks, the team could control the ability of the fish to bounce polarized light off their scales. Without polarization, the males lost the attention of the females³.

This was not the first time that Cummings had predicted that male ornamentation is based on beholder perception. Earlier in her career, she dove about nine metres below the surface of the ocean in the Californian kelp forests, where

“You can’t look at these frogs, and not think that something is going on.”

light becomes fluid and patchy. Seeing the variety in lighting across the forest, Cummings suggested that surfperch fish see and display features that are tailored to their specific forest backdrop.

When she examined the eyes of one surfperch species, *Hypsurus caryi*, that swam through variously lit parts of the forest, she found an abundance of opsin proteins⁴. The spectral bands of light that each opsin absorbed hardly overlapped that of other opsins — a trait that enables the fish to perceive greater variation in colours. Males from this species have colouring that corresponds to this visual bias: they have blue and orange markings that stand out against the greenish hue of the algae-filled water. Meanwhile, the surfperch *Embiotoca lateralis*, which dwells in the densest and dimmest regions of the kelp forest have different eye anatomies. Their opsin proteins cover overlapping spectral regions. These surfperch easily sense the contrast between light and dark, but they are less sensitive to differences in hue. As a result, the allure of the males of this species depends on illumination. Thousands of years ago, these two surfperch species shared a common ancestor. Cummings speculates that their descent from that ancestor might have begun with adaptations that helped the fish to see in their distinct environments. Over time, females developed preferences for males that exhibit features that they could easily see.

Cichlid fish off the Tanzanian shore of Lake Victoria seem to be evolving in this manner as their populations stop mating with one another in the wild — the first step in speciation. At this point, some populations are separate enough from one another that they can be considered distinct species, even though they will breed if isolated in captivity. The fish even differ in the type of visual proteins that they use to perceive colours. Martine Maan, an evolutionary biologist at the University of Groningen in the Netherlands, and her colleagues have found that cichlids in the shallows,

which include a broad range of solar wavelengths, perceive a wide spectrum of colour. In turn, males commonly display blue designs, which Mann suggests might hide them from avian predators, while revealing them to female cichlids swimming nearby. Deeper down, at depths birds cannot see and where red light is predominant, photoreceptors in cichlid eyes are shifted towards the longer wavelengths — and the males are redder⁵.

Although the two populations of cichlids are neighbours, they no longer mate in the wild because females prefer males that sport the colours that they see best. They might lose the physical ability to interbreed as the populations diverge further. “I’m trying to figure out if adaptations in the visual system, which are driven by ecological requirements, have consequences in how females perceive male colours,” says Mann. “That would provide a fast route to speciation.”

SHOWING OFF

These studies do not explain why female fish are so taken by a wash of blue, or why the female butterfly is so dazzled by bursts of light. Their preferences are determined by what their senses tune into — but what do the traits mean?

Occasionally, decorations link directly to benefits. For example, black swans with curly tail feathers tend to be preferred by the opposite sex, and they often occupy the most territory. And, while peering at a common fruit fly at the University of Tromsø in Norway, entomologist Jostein Kjærandsen, discovered a form of beauty that Darwin never suspected, and that seems to come with a pay-off. He noticed that the *Drosophila melanogaster*’s wings reflected a purplish hue against a black background. The wings of other flies from the same species reflected different colours. Soon after, he and his colleagues demonstrated that female fruit flies



The common yellowthroat uses its yellow feathers or black mask to attract mates.



Brightly coloured butterflies tend to be more resilient individuals.

mated more often with males that reflected magenta, as opposed to yellow or blue, sheens⁶. The colours varied depending on the thickness of the wing, the team found, prompting the researchers to speculate that the sheen subtly indicates how well the wings allow flies to control flight. That is a genetically controlled trait that females would find advantageous to endow to their offspring, says Kjærandsen's colleague Erik Svensson at Lund University in Sweden.

D. melanogaster is particularly amenable to advanced genetic manipulation, giving Svensson the opportunity to test the hypothesis. "If we can identify one or several genes that alter characteristics of wings, we could use gene-silencing techniques to manipulate those characteristics and look at the effect on female choice," Svensson says.

Just as frequently, however, beauty links to no obvious benefit. In 1975, biologist Amotz Zahavi proposed the handicap hypothesis to account for extravagant characteristics that impair their bearers, but ironically attract mates⁷. Zahavi highlighted the peacock's blue and gold tail that was so loathed by Darwin. It made the birds easy for predators to spot, and gave parasites plenty of feathers to attach to. But the very fact that it persists through the generations means that females like the tails, possibly because the feathers identify males that are healthy enough to withstand the negative effects. Likewise, Kemp suggested that butterflies with the brightest ultraviolet markings make for easier prey than their duller counterparts, but they also withstand turbulence better⁸. Indeed, he found that brilliantly winged male butterflies survived flashes of hot and cold, and malnutrition during juvenile stages more

"A woman might find a guy with huge muscles attractive, but I might find him intimidating."

often than did their less-flashy counterparts. Decorations on wings do not themselves confer resilience, but Kemp's study suggests that the trait reveals that vital, but invisible quality. "Females get a genuine glimpse into the potential quality of their mate's genetic quality simply by appraising the quality of his iridescent signal," he says.

SIGNALLING FITNESS

Earlier this year, researchers found support for the good-genes theory in common yellowthroat warblers, *Geothlypis trichas*. In New York, female warblers prefer males with large, bright yellow breast feathers. But around 1,500 kilometres west, in Wisconsin, females rate males on the size of their so-called masks — black feathers around their eyes. Despite the different preferences, however, the quality — size and colour — of both yellow bibs and black masks indicates the power of the individual bird's immune system. Greater variation in the genes that are essential in immune responses, called the major histocompatibility complex or MHC genes, correlate with better bibs or masks; this in turn enhances the bird's ability to fend off diverse infections⁹. Females' preference for a bib or a mask is rather arbitrary: it is the signal the features send that counts.

Furthermore, a certain style — a bib, a mask or a haircut, for example — can send different signals, depending on who is looking. "A woman might find a guy with huge muscles attractive, but I might find a guy with huge muscles intimidating," says Ken Kraaijeveld, an evolutionary biologist at VU University Amsterdam. That is efficient from an evolutionary perspective — better to use existing features than develop a feature anew. Kraaijeveld warns that dual-function features can obscure biologists' view. If researchers are most interested in sexual selection, he says, they might focus on how a male bird

behaves to attract females in the mating season, and neglect to observe how that same behaviour helps them to find food in the winter.

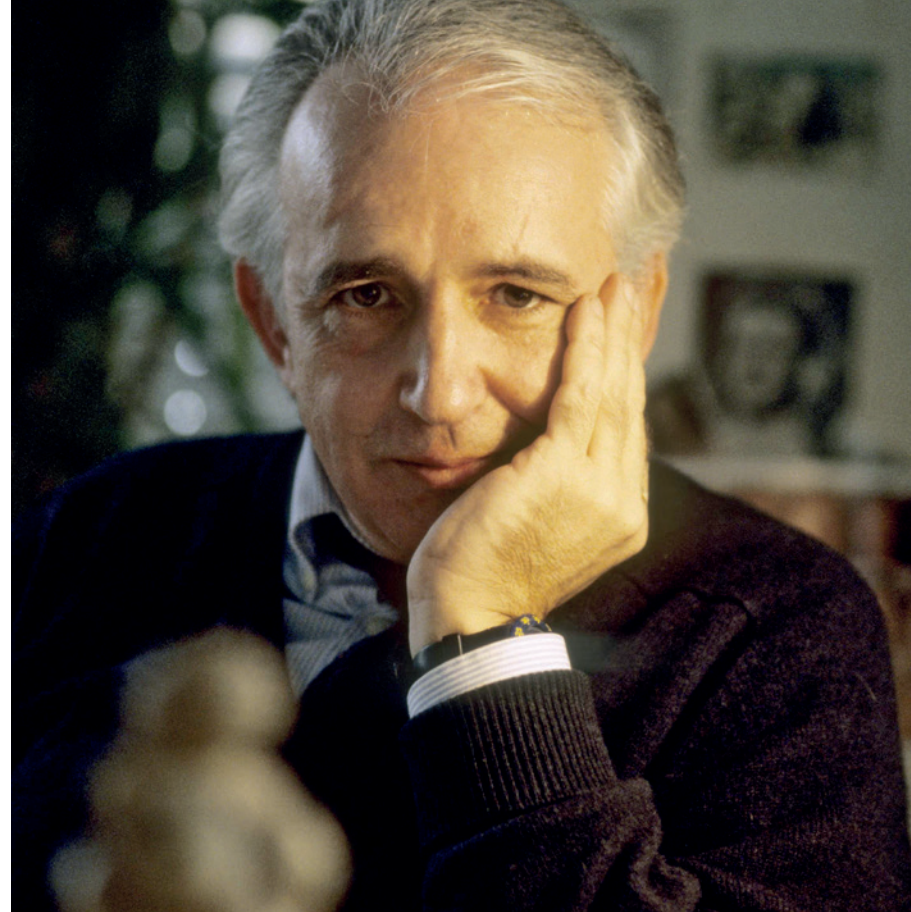
To complicate matters further, not all males and females want to say 'come mate with me'. Cummings is now comparing male behaviour of various swordtail species. "One type of fish includes the knights of the species, they shimmy to attract females, and if she gives them the right cue, they have cooperative copulation," Cummings says. Not surprisingly, these flirty males are colourful. Meanwhile, males from a closely related species mate by simply thrusting their fish penis, a gonopodium, into an unsuspecting female. Not only are these crude males undecorated, but Cummings has found that females in this species are wired differently than courted females. Specifically, more neural genes associated with learning and memory are activated in courted female fish when they interact with males¹⁰. That is a tantalizing finding because it suggests that a female's preference may be acquired, and not just genetically determined. In this way, the beholder of beauty can speed up evolution's trajectory.

Although she cannot yet explain it, Rayna Bell is confident the beauty she saw in African reed frogs has meaning — she just needs to discover what the creatures are seeing and silently saying. Few biologists have been to those islands off Africa's west coast. "You don't just swoop in and get the sexy story," Bell says. "That takes time, but I don't mind. Its exciting to start from almost zero, to realize there's so much diversity that we know so little about."

Cummings, for her part, sees meaning in the 'flush of a maiden's cheek' that Darwin found haphazard. It is no accident that cosmetic companies sell pink rouge and red lipstick, rather than blue. "Red mimics a youthful glow," she says. "It advertises that the wearer is young and reproductively valuable." And from a survival of the species perspective, that is truly a beautiful thing. ■

Amy Maxmen is a freelance science writer in Berkeley, California.

1. Darwin, C. R. *The Descent of Man, and Selection in Relation to Sex* (John Murray, 1871).
2. White, T. E., Zeil, J. & Kemp, D. J. *Evolution* **69**, 14–25 (2015).
3. Calabrese, G. M., Brady, P. C., Gruev, V. & Cummings, M. E. *Proc. Natl Acad. Sci. USA* **111**, 13397–13402 (2014).
4. Cummings, M. E. *Evolution* **61**, 530–545 (2007).
5. Seehausen, O. *Nature* **455**, 620–626 (2008).
6. Katayama, N., Abbott, J. K., Kjærandsen, J., Takahashi, Y. & Svensson, E. I. *Proc. Natl Acad. Sci. USA* **111**, 15144–15148 (2014).
7. Zahavi, A. *J. Theor. Biol.* **53**, 205–214 (1975).
8. Kemp, D. J. & Rutowski, R. L. *Evolution* **61**, 168–183 (2007).
9. Whittingham, L. A., Freeman-Gallant, C. R., Taff, C. C. & Dunn, P. O. *Mol. Ecol.* **24**, 1584–1595 (2015).
10. Cummings, M. E. *Anim. Behav.* **103**, 249–258 (2015).



Q&A Karl Grammer

Innate attractions

Karl Grammer, professor of anthropology at the University of Vienna, has been a pioneer in human attraction and courtship research. He discusses what he and others have learned by studying human beauty from an evolutionary perspective.

Why do you believe that our perceptions of human beauty were shaped by evolution?

In all other animals, appearance plays a big part in mate selection and reproductive capability. I am a biologist, so I believe that this cannot be different for humans. Humans are obsessed with beauty. Beautiful children get less punishment than less-attractive children for the same misbehaviour. Even babies look more frequently at beautiful faces. When you find an obsession like this, there must be something deeper than a simple cultural norm. There are 3,000-year-old poems that talk about beauty and love — so this obsession goes through the whole history of mankind.

So you disagree with those who argue that standards of beauty are culture-bound?

Yes. People always say that beauty standards are generated, for instance, by fashion models. I do not think that is true. Models might have some influence, but only on a very small scale. Some argue that beauty is a myth — that “real beauty comes from inside”. This is completely untrue. Beauty provides reliable information about youth, fertility and health.

What is the evidence that human beauty is an indication of Darwinian sexual selection?

Beautiful people are healthier than less-attractive people — this has been shown repeatedly. We have shown that more-attractive women produce more offspring over a lifetime than do less-attractive women. Studies also find remarkable consistency in the facial and bodily features that people find beautiful — even across different cultures, races and ages. In one of our studies, people from South Africa and Austria judged the same Japanese women to be attractive. You would not expect this unless there was a biological basis behind beauty.

Which facial and bodily features are consistently judged as attractive?

We and others have identified eight pillars of beauty: youthfulness, symmetry, averageness, sex-hormone markers, body odour, motion, skin complexion and hair texture. I think this line of research is almost finished. It is no longer useful to just decide that something is more beautiful than something else. The signals of beauty have been identified. The next step is to try to work out what these signals

are for. For example, symmetry is thought to reflect stable development and parasite resistance, and body odour is thought to convey information about the immune system. But direct evidence for these connections is weak. Also, specific genes involved in determining attractiveness have yet to be pinpointed.

What do you consider your most important discovery?

We have shown that beauty signals are redundant — they tend to go together. If you have a nice face, your body odour smells good; if you smell good, you are more symmetrical; if you have a nice voice, you have a nice face; and so on. The whole body is one ornament; it is not just an array of independent signals. For us, this makes it highly likely that there is a biological basis. This also means that to uncover the connections between the cues and the underlying biology, researchers need to study multiple features simultaneously rather than one at a time.

If beauty standards are innate, why is there cultural variability?

We have to be able to adjust our beauty standards to the mean of the population we are living in, or we would run the risk of never finding a mate. The eight pillars of beauty are construction rules. As long as you adhere to the construction rules — such as averageness and symmetry — then the specific content can vary. Attractiveness has to be a flexible concept to increase the pool of potential mates.

With plastic surgery and cosmetics we can artificially manipulate beauty. Won't this remove, or at least reduce, the selection pressures on beauty?

For a long time in human history beauty was a non-falsifiable signal. Now it is falsifiable. We do not yet know the consequences of this because too little time has passed; we would need another 10 or 20 generations of plastic surgery to see the evolutionary effects. One thing that may make beauty harder to falsify is that the signals are redundant. So you might be able to change your facial symmetry, for instance, but not your body odour.

Has human-attraction research ever been considered objectionable?

Yes. You will not find many publications on body or facial appearance from the 1960s or 1970s. It was considered politically incorrect at that time to judge people on their appearance. When we started our work in the early 1990s, we did not worry about this. We were biologists, and we knew that symmetrical scorpion flies attract more mates than asymmetrical ones. We believed that what applies to the scorpion fly also applies to humans. And that is how the whole thing took off. ■

INTERVIEW BY KRISTIN LYNN SAINANI

This interview has been edited for length and clarity.



Faced with images of unrealistic ideals, men are increasingly concerned about their appearance.

MASCULINITY

Men's makeover

Historically, women have been the focus of body-image studies. But as men pay more attention to their appearance, researchers are forming a clearer picture of male self-image.

BY KELLY RAE CHI

Insecurities about body shape and size are a frustratingly common topic of conversation among groups of women and girls. Body-image research has shown that participating in, or even just hearing, such 'fat talk' fuels appearance dissatisfaction in women.

For the past few years, whenever Northwestern University psychologist Renee Engeln presented these results, audience members would ask, 'What about men? Do men do this too?' she recalls. Intrigued by this question, she and her colleagues, based in Evanston, Illinois, designed a fat-talk scale for men. They found that men do it, too, but only in specific contexts¹. 'Men talk about body dissatisfaction when they're eating and when they're at the gym,' says Engeln. 'Women talk about body dissatisfaction when they're talking.'

Feeling bad about one's body is among the strongest predictors for developing an eating

disorder, and one of the most modifiable. Interventions aimed at addressing such concerns are better studied in women, who are more likely than men to have a recognizable eating disorder and who have been subject to more of the superhuman beauty ideals that pervade the media. Over the past decade, however, boys and men have been exposed to similarly unattainable standards.

The evidence is in the aisles. Superhero costumes for boys feature chiselled abs, and health and beauty products for men line shop shelves. Sales of men's grooming products have skyrocketed across the globe over the past few years. 'Men are being addressed as consumers of health and beauty products and services in a very targeted way, in ways they haven't been historically,' says Brendan Gough who studies men's body-image issues and masculinity at Leeds Beckett University, UK. According to Gough, some young men are thought to be injecting the oil synthol

into their muscles to make them look larger or taking diet pills that contain the appetite suppressant ephedrine to lose weight. Body dissatisfaction can become an obsession and can lead to clinical disorders (see page S14). These negative feelings can also trigger symptoms of depression.

Research is starting to examine how men feel about their bodies, particularly when faced with images of masculinity in the media. A better understanding of how boys and men deal with insecurities about their appearance will help with the design of initiatives that are aimed at preventing unhealthy behaviours such as eating disorders. And, as studies suggest, this cannot come soon enough.

EVOLVING MANSCAPE

Psychological and social-sciences research into male body image has been around only for the past 15 years or so, says Philippa Die-drichs, a health psychologist at the University of the West of England in Bristol, UK. But epidemiological studies are now starting to address the prevalence of concerns.

An almost two-decade-long study of US teenagers — the Growing Up Today Study, or GUTS — reported that nearly 18% of adolescent boys are highly concerned about their physical appearance². Of the male respondents, 7.6% reported using muscle-building supplements, growth-hormone derivatives or anabolic steroids to achieve their ideal body. Although these behaviours do not fulfil the conventional eating-disorder criteria, they are risky actions that may be missed by paediatricians and parents, says Alison Field, an epidemiologist at Boston Children's Hospital in Massachusetts and one of the study's authors.

When GUTS began in 1996, research questionnaires for assessing body image in boys and men were not available. 'Most of the large studies ask the same questions of males and females, with the assumption that if someone is concerned about their weight it looks the same,' she says. The story is similar for Lina Ricciardelli, a psychologist at Deakin University in Victoria, Australia, who studies body-image issues. Only 20 years ago, she says, the prevailing wisdom was that men do not have these problems.

UNDER THE INFLUENCE

A host of factors have been tied to body-image issues in males. Men who are not dating, for example, tend to be more affected by media exposure than those who are. Sexual orientation plays a part as well. Compared with heterosexual men, gay men are more likely to express dissatisfaction with their appearance and are at greater risk of developing an eating disorder. The focus on appearance is ingrained in gay culture; gay media emphasizes unrealistically muscular and lean models to sell beauty products to an expanding and

TIM TADDER/CORBIS

powerful consumer base.

The mainstream media exerts a powerful influence on body-image perception. For instance, one study found that male students in the United States viewing 30 minutes of television with commercials that feature muscle-bound men were more likely to report feeling depressed and dissatisfied with their own bodies compared with participants who watched television with neutral advertisements³. And a meta-analysis of 25 studies demonstrated a link between media exposure and measures of body dissatisfaction, low self-esteem and depression⁴. “The people more at risk are the ones who believe in the messages of the media,” says Ricciardelli — they internalize the values promoted by television and magazines, believing them to reflect reality.

The way that men manage their masculinity is important for understanding and addressing their body-image concerns, says psychologist Viren Swami of the University of Westminster in London. His work centres on the idea that when some men feel their masculinity is threatened — for instance, by the idea of gender equality or more-discrete events, such as being turned down for a date — they try to reassert some of that masculinity in the gym. He is in the early stages of examining whether the promotion of more-egalitarian attitudes between men helps to address their body-image concerns.

Although what is considered masculine behaviour has become more flexible of late (for example, new fathers now take on more domestic duties than before), men often do not consider grooming and dieting to be part of a traditional masculine role, says Gough. Instead, men hoping to lose weight say they want to become fit and strong. And straight men who wear make-up say that they use it because it will help them to attract women or be more successful at their jobs⁵.

MEASURES FOR MEN

One common way to measure body-image dissatisfaction is to show a study participant drawings of variously sized figures, and then to ask him to pick out

drawings that represent his current body size and his ideal size. The bigger the difference between the two, the more dissatisfied a person is with his body. Software has lent more precision to this measure by allowing people to see themselves on a projected screen and to adjust their bodies to their ideal, says psychologist Rick Gardner, a body-image perception specialist at the University of Colorado in Denver.

Body image is more than just a one-off result. It is constantly changing, not as a result of any one individual characteristic, but with changing cultural and societal pressures, says Glen Jankowski, who is part of Gough's group at Leeds Beckett.

Even as masculinity norms become more flexible, men may feel less able than women to discuss body-image and emotional issues. However, in research at least, male participants do willingly talk about body-image concerns, says Engeln. At the outset of her research, she says, “we had some concerns that even if men did engage in something like fat talk, they wouldn't openly admit it. But our initial survey results showed us that men were quite open to talking about their body concerns, and many readily admitted having body-focused conversations with other men.”

ADAPTING THE INTERVENTION

Body image is realized surprisingly early in life. Children as young as 5 years can develop a negative body image, according to Ricciardelli. That is why it is important to reach them while they are young, she says. She and others, including Swami, are developing programmes that promote an appreciation of what the human body can do, for example through physical activities such as contemporary dance.

Body-image programmes for boys are based on less-solid scientific foundations than those for girls, says Diedrichs. There is no single standout programme that has been consistently shown to be effective for boys, she adds.

Diedrichs, Jankowski and their colleagues are adapting a

“There is a whole industry out there that does not want you to feel good about yourself.”

popular and effective intervention for women — the Succeed Body Image Programme, known in the United States as the Body Project — for use in men. So far, the data look promising, Jankowski says: at 3-month follow-ups, the men who took part in 2-day workshops felt better about their bodies and were less worried about their muscularity and body fat. Less clear from the researchers' initial studies, however, was the effectiveness of this programme when it was adapted for adolescent boys.

In developing an approach for Israeli teenagers, Moria Golan of the Hebrew University of Jerusalem knew that the US and UK discussion-based programmes would not suffice, in part because Israelis already openly discuss body image. Her group developed the In Favor of Myself wellness programme, which consists of 8 sessions, each lasting 90 minutes, that use games and interaction to build self-esteem through qualities other than appearance.

In a study of children aged 12 to 14 years, Golan and her colleagues found that girls gained more from the programme than boys⁶. The researchers have since developed a boy-friendly programme and are now testing it, with mixed results. Whether the group is male only or mixed gender, differences in the children's ethnicities, and whether the intervention is delivered by a teacher, or an outsider can all factor in a programme's effectiveness, she says.

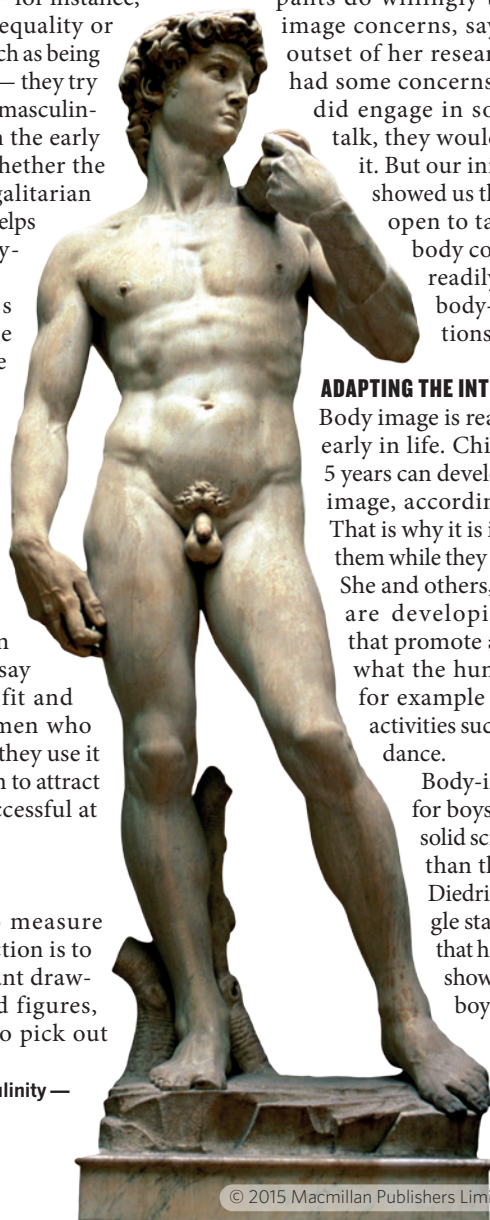
Body-image concerns are closely tied to cultural and societal pressures. Young white men living in North American, European or Australian cities have been the focus of most studies, but what researchers learn from these men may not apply to all ethnic groups, or to boys and men living in developing countries who are increasingly exposed to Western media and ideals. Research, however, is expanding to include these groups.

In the meantime, thinking bigger than group interventions and trying to effect change in a society that promotes unrealistic ideals can be overwhelming. “We need to help young people become more savvy that they're heavily marketed to,” Field says. “There's a whole industry out there that does not want you to feel good about yourself.” Women have been coming to grips with this for generations. Now men are learning to deal with this manufactured inadequacy. ■

Kelly Rae Chi is a freelance science writer based in Cary, North Carolina.

- Engeln, R., Sladek, M. R. & Waldron, H. *Body Image* **10**, 300–308 (2013).
- Field, A. E. et al. *JAMA Pediatr.* **168**, 34–39 (2014).
- Agliata, D. & Tantleff-Dunn, S. *J. Social Clin. Psychol.* **23**, 7–22 (2004).
- Barlett, C. P., Vowels, C. L. & Saucier, D. A. *J. Social Clin. Psychol.* **27**, 279–310 (2008).
- Hall, M., Gough, B. & Seymour-Smith, S. *J. Men's Stud.* **20**, 209–226 (2012).
- Golan, M., Hagay, N. & Tamir, S. *PLoS ONE* **9**, e91778 (2014).

ROGER ANTRUBUS/GETTY IMAGES



A classic example of masculinity — Michelangelo's *David*.



combination of antidepressants (typically at high dosages) and talk therapy. However, BDD is not simply a clinical variant of OCD, and in the past few years researchers have begun to explore ways to tailor treatments to specifically address people's excessive concerns over their appearance.

Some targeted forms of psychotherapy are the focus of randomized controlled trials, and researchers are scanning patients' brains to learn more about how to correct the neural circuitry that is responsible for BDD. "We can now offer empirically based treatments that often work," says Katharine Phillips, a psychiatrist at Alpert Medical School of Brown University in Providence, and author of *Understanding Body Dysmorphic Disorder* (Oxford Univ. Press, 2009).

Relief cannot come soon enough for Jessica and others with the condition. "It's very painful to have this disorder," says Sabine Wilhelm, a psychologist at Massachusetts General Hospital. "Some patients are so sick that they're almost completely housebound."

COSMETIC CONCERNS

BDD manifests in many ways. One person might think his eyebrows are uneven or that his muscles are too small. Another obsesses over her pointy chin or acne scars. "Any body part can be the focus of concern," says Wilhelm.

Consumed by their imagined ugliness, people with this condition often have severe depression, and engage in substance misuse and life-threatening behaviour. According to data compiled by Phillips and her colleagues, the suicide rate of those with BDD is at least 22 times greater than that of the general population — making BDD one of the most lethal psychiatric conditions.

The disorder affects around 2% of the overall population, and yet most cases go unrecognized and untreated. Instead of seeking the help of mental-health counsellors, many people with BDD visit cosmetic surgeons, dermatologists and dentists. Most patients who have appearance-enhancing procedures, however, simply shift the focus of their concerns or they continue to worry about imperfections in the treated area.

Lisa Ishii is a plastic surgeon at Johns Hopkins School of Medicine in Baltimore, Maryland, who is calling on physicians in her field not to operate on people with BDD. "They don't need cosmetic surgery," she says. "They need psychiatric care." Ishii and her team have begun to use a two-stage screening process — a questionnaire followed by a clinical interview — to distinguish patients with BDD from those who are merely dissatisfied with certain physical traits¹. This approach not only helps people to find the right type of care, Ishii says, but it also protects the interests of plastic surgeons — some of whom have been sued, physically threatened or even killed by dissatisfied people with BDD².

Still, many surgeons are reluctant to implement such a screening instrument because they

MENTAL HEALTH

Monsters in the mirror

Researchers are probing how brain circuitry goes awry in people with body dysmorphia and how to treat the condition.

BY ELIE DOLGIN

Jessica's body-image problems started early. In middle school, it was the frizziness of her hair. In high school, it was the size of her nose. Then last year, during law school in Massachusetts, Jessica's insecurities about her looks ballooned into a full-blown fixation. At the age of 25, she began to worry non-stop about the smallest signs of ageing. (At Jessica's request, we are using only her first name.)

She hated her hands, which she saw as blotchy and venous. She thought the skin on her face was thin and wrinkly. She would search for grey hairs and pluck them out of her head. Obsessing over these features occupied Jessica's thoughts for up to ten hours each day. "I had all these doomsday ideas about what my appearance

meant for my future," she says. "I had this fundamental belief that if I didn't look like a fresh ingénue that no one would give me a chance."

Then, last year, Jessica saw an advertisement on the subway that would change her life. Psychiatrists at the Massachusetts General Hospital in Boston and the Rhode Island Hospital in Providence were looking for study participants with a condition called body dysmorphic disorder (BDD). This severe mental illness is characterized by chronic, often delusional, pre-occupations with non-existent or slight flaws in appearance that extend far beyond vanity. That is when it clicked for Jessica. Maybe her problems were not physical, but psychological.

BDD shares a number of features with obsessive-compulsive disorder (OCD), and it is often managed in much the same way — through a

believe that their intuition serves them well enough. “And therein lies the problem,” Ishii says. She has unpublished survey data showing that most cosmetic surgeons think that they can pick up whether a patient has BDD without psychiatrically validated scales and measures. “But actually,” says Ishii, “most can’t.”

FORCED EXPOSURE

The trial that Jessica discovered is run by Phillips and Wilhelm and is testing whether a treatment strategy known as cognitive behavioural therapy (CBT) is more effective than supportive psychotherapy at helping people with BDD cope to with and overcome the disorder. Jessica was randomly assigned to the CBT group. In February 2015, she attended her first therapy session. For the next six months, she learned new skills to challenge and sidestep negative thoughts whenever they arose.

Part of CBT involves exposing people to the thoughts and situations that create intense anxiety for them. Patients then learn how to face the anxiety without engaging in the behaviours that reinforce and maintain their symptoms. For Jessica, this meant going out in public without makeup — something she had not done since her university days. At first, Jessica says, “it felt like being sort of naked.” But thanks to the coping tools she learned for dealing with unhelpful patterns of thinking and behaviour, Jessica often goes entire days without her cosmetic defences. “CBT has been tremendously successful for me,” she says. “There are certain scenarios where these cycles of negative thoughts will increase in frequency. But now I have the skills to put the kibosh on them.”

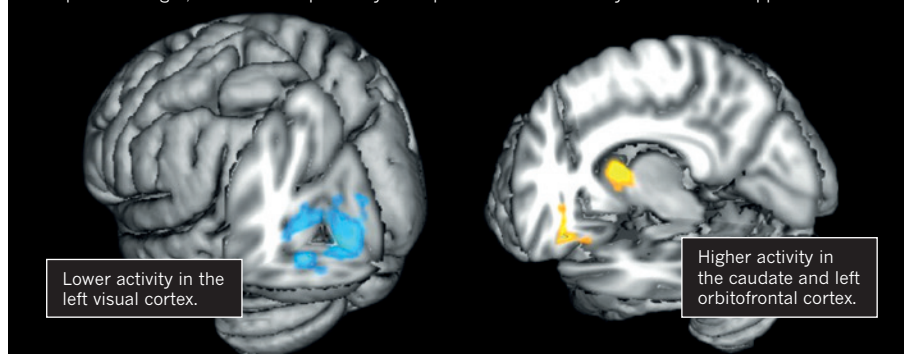
Jessica is not alone. In an earlier trial at Massachusetts General and Butler Hospital in Providence, Phillips, Wilhelm and their team found that 50% of participants showed improvements in their symptoms after 12 weeks of CBT compared with 12% of those who did not undergo therapy⁷. Everyone in the study then received a 22-session course of CBT and 24 of 29 people who finished the trial responded favourably. To test whether this dramatic response rate was thanks to the specifics of CBT and not just the therapeutic experience more broadly, Wilhelm and Phillips set up the larger, randomized trial of 120 participants that Jessica took part in.

These are two of a number of trials the researchers have been involved in. Phillips recently presented the results of a study of relapse rates among people with BDD after they stop taking antidepressants. And Wilhelm is running a 50-person, placebo-controlled trial to test whether a neurotransmitter-activating drug called D-cycloserine can enhance the behavioural learning that happens during CBT — a strategy that has worked in the treatment of anxiety disorders such as OCD.

But aside from the few clinical centres that specialize in BDD, CBT is not widely available. And even when the therapy is an option, many patients feel too ashamed to openly discuss their

BRAIN IMAGING

Imaging techniques have shown that people with body dysmorphic disorder have unusual brain activity in areas that process images, which could explain why these patients obsess over tiny defects in their appearance.



problems with a therapist. Psychiatrist Christian Rück and his colleagues at the Karolinska Institute in Stockholm hope to overcome these obstacles by delivering CBT over the Internet — a practice known as iCBT.

In a 12-week pilot study of iCBT, they found that 18 of 22 patients responded to therapy⁴. At the first International Conference on BDD in London in May, Rück's team presented impressive follow-up results. In a 94-person, randomized trial, iCBT outperformed supportive psychotherapy. The web-based protocol still requires therapist involvement through a built-in e-mail system, but each mental-health professional spends, on average, about 10 minutes with a patient each week, instead of the usual 45–50 minutes. With iCBT, “there might be one or a few people in Sweden who could treat a whole nation”, Rück says.

“They don’t need surgery, they need psychiatric care.”

BRAIN RETRAINING

To explain the biological basis of BDD and responses to therapy, many scientists have turned to neuroimaging. Psychiatrist Jamie Feusner at the University of California, Los Angeles, and his colleagues have shown that connectivity patterns between brain regions in people with BDD are different from those of individuals without body-image problems⁵ — and that brain activity is particularly abnormal in areas that are responsible for processing visual stimuli⁶ (see ‘Brain imaging’).

This irregular visual system in the brain could explain why people with BDD tend to obsess over minute body details, but miss the bigger picture. To help rewire the brain, Feusner is testing a type of perceptual retraining that involves activities designed to help people adjust their visual balance from detail-oriented to global processing. One such exercise attempts to modulate eye gaze by asking individuals to view a digital photograph of their face and then hold their visual focus within a target circle between the eyes (instead of on, say, a barely visible facial scar). Another presents

the same picture but for only a split-second, forcing the brain to process the face more holistically. If such interventions lessen symptoms of this disorder, Feusner says, they would be “the first to be directly informed by knowledge of aberrant neurobiology in BDD.”

Phillips is also probing the genetics of BDD in search of new drug targets. In collaboration with a team at the University of Toronto in Canada, she identified a gene that encodes a brain receptor involved in the transport of the neurotransmitter γ-aminobutyric acid (GABA) that may be implicated in the development of the disorder⁷. The researchers are now engineering mice with mutations in this gene to create the first BDD-specific animal model. They plan to assess how early life stressors in these mice affect the development of grooming behaviours (people with BDD are commonly preoccupied with grooming). Eventually, they hope to test which drugs offer symptom relief too.

As for Jessica, her treatment was so successful that she rarely dwells on her appearance for more than about 30 minutes per day — a massive reduction from the 10 hours she was spending. Before finding psychiatric help, Jessica had met with a dermatologist, who had recommended laser surgery to remove a series of tiny bumps called syringomas that had developed under her eyes. She had all but committed to go through with this surgery. But, she says, following her therapy, “I’ve decided not to do it.” Thanks to CBT, those bumps are “no longer these outsized disfigurements that I once perceived them to be,” Jessica says. “They’re so small, it doesn’t really matter.” ■

Elie Dolgin is a science writer in Somerville, Massachusetts.

1. Dey, J. K. *et al.* *JAMA Facial Plast. Surg.* **17**, 137–143 (2015).
2. Sarwer, D. B. *Aesthet. Surg. J.* **22**, 531–535 (2002).
3. Wilhelm, S. *et al.* *Behav. Ther.* **45**, 314–327 (2014).
4. Enander, J. *et al.* *BMJ Open* **4**, e005923 (2014).
5. Arienzo, D. *et al.* *Neuropsychopharmacology* **38**, 1130–1139 (2013).
6. Feusner, J. D. *et al.* *Arch. Gen. Psychiatry* **67**, 197–205 (2010).
7. Phillips, K. A. *et al.* *J. Obsessive Compuls. Relat. Disord.* **6**, 72–76 (2015).



Q&A David Deutsch

Objective beauty

Physicist David Deutsch is considered the founding father of quantum computing. In his 2011 book, *The Beginning of Infinity*, Deutsch argues that there is such a thing as objective beauty.

What is your argument for the existence of objective beauty?

The argument I like best is about why flowers are beautiful. Flowers evolved to attract insects, and insects evolved to be attracted to flowers. But this explanation leaves a massive gap: it only explains why insects like flowers. So how is it possible that something that evolved to attract insects can be attractive to humans too? I conclude that there must be objective beauty — aspects of beauty exist outside cultural fads or sexual selection. And these aesthetic truths are as objective as the laws of physics or maths.

If beauty is objective, why is there so much variation in what people consider beautiful?

Beauty has both a subjective and objective part. Human aesthetic judgment is a complicated mixture of genetic, cultural and objective factors. If you look at paintings from centuries ago, you will find that the women tend to be considerably heavier than what we now consider to be ideal. That can be neither objective nor genetic, so it must be cultural. Our preference for symmetry is probably related to our preference for healthy mates — many diseases and deformities make people less symmetrical. So that one could be genetic.

Our knowledge of the nature of objective beauty is still primitive. We cannot reliably distinguish between subjective and objective beauty, certainly not by just looking. Things that

meet aesthetic preferences built into our brains or instilled by culture look just as beautiful to us as those that are objectively beautiful.

Why is it important to acknowledge the existence of objective beauty?

During the twentieth century, some movements denied that there was such a thing as objective truth in science. These movements significantly held back scientific progress. For example, I'm pretty sure quantum computing would have been proposed in the 1950s rather than in the 1980s if it had not been for these beliefs. Because our culture generally denies the existence of objective beauty, research into it is substantially cut down. I'm not aware of any research that looks at the nature of objective beauty.

How do you counter those who insist that beauty is always subjective?

It is remarkable how the arguments against objectivity in aesthetics, and in morality, have exact counterparts in classic arguments against objectivity in science. People say we do not have access to the world; we only have access to the interpretations that we put on the world through our senses. The second part is right, but that does not mean we cannot achieve truth. To think that, is to confuse truth itself with some sort of superhuman, certified, reliable access to the truth. For example, the abolition of slavery was an objective moral improvement. It is not just cultural. It is certainly not genetic. It is not

a matter of preference. It would still be true that slavery was wrong even if nobody knew that.

What is the connection between aesthetic beauty and scientific argument?

Beauty in science is called elegance. Physicists will, as a matter of practice, take elegance as a guide. There is the phrase: many a beautiful theory was slain by an ugly fact. This is very true. But when it happens, we inevitably find an underlying theory that is even more beautiful than the theory that was slain. So beauty cannot be used as a criterion of what is true; but it is at the very least useful as a guide to what to try next.

What factors do you believe govern human sexual attraction?

I speculate that human beauty started out just like any other animal beauty — completely biological, and not objective at all. But as humans became intelligent and started making aesthetic judgements, they increasingly tried to improve the aesthetic and other standards by which they chose their mates. And that increasingly led to true standards. So we should find that the common features that have changed in all human populations since our ape ancestors are aspects in which humans have become objectively more beautiful.

Are you saying that humans have steadily made the world more beautiful in the same way that we have achieved scientific progress?

Yes. Objective beauty, like objective truth, is subject to open-ended improvement. For example, our knowledge of physics can contain more and more truth, even though no one theory is ever perfectly true. Newton's theory contained more truth than what was there before. But it was superseded by Einstein's theory. And science continues its progress by finding new aspects of reality forever. By contrast, something that is subjective reaches a maximum and then stops.

We discover aesthetic truths in the same way as we discover scientific truths, even if the methods look different. It is conjecture and improvement according to some standard; then improvement of the very standards; then criticism of existing ideas according to these standards; and so on.

Aesthetic progress has been a lot slower than scientific progress because people can only express in words a tiny proportion of what they know about beauty. But humans have achieved an enormous amount. Mozart and Beethoven improved artistic standards in music. And films have become more beautiful in the past century.

Only humans can improve on beauty. When nature achieves beauty it is an accidental by-product of something else. Nature can only get so beautiful, but humans can paint something that is more beautiful than any scene. ■

INTERVIEW BY KRISTIN LYNN SAINANI

This interview has been edited for length and clarity.



Scientists are fascinated by the biological, social and medical implications of beauty. Here are four of their most pressing questions.

BY CHELSEA WALD

QUESTION

WHY IT MATTERS

WHAT WE KNOW

NEXT STEPS

1

What is the point of human beauty?

Beauty is hard to define, but we know it when we see it. Although today this superficiality often seems pointless and even destructive, it may have served a useful purpose for our distant ancestors.

Some traits that humans find beautiful may correlate with health and reproductive viability, but preferences for certain traits could simply have evolved as by-products of the way in which the brain processes information.

Models that incorporate dozens of variables for describing facial features are better ways to test evolutionary and non-evolutionary hypotheses, and may offer insight into how we weight different cues in human faces.

2

How can we overcome our obsession with physical beauty?

Many researchers think that attractiveness is too closely bound up with personal worth in society. This can lead to prejudice, as well as psychological conditions such as eating disorders and depression.

When we find others attractive, we tend to assume that they are also good people. Being attractive can lead to benefits at work, in the courtroom and in politics. Media images and 'fat talk' can fuel negative body images.

Interventions that treat body-image disorders are having some success, but they need to be adapted for men and all ethnic groups. Work on other types of prejudice could open up ways to override our subconscious beauty bias.

3

What is so special about youthful skin?

Some people retain young-looking skin as they age, whereas others resort to creams and procedures. Little is known about the mechanisms of skin ageing, including whether younger-looking skin — however it is achieved — is any healthier.

Studies have revealed many gene variants and molecular pathways associated with skin ageing. Combining those findings with data on proven treatments, such as topical creams and broadband light therapy, could tease out the biochemical details of skin ageing.

Cosmetics companies hope to use these findings to develop products that will delay or reverse skin ageing, but further steps are needed before any personalized, evidence-based treatment comes to market.

4

Why do we take pleasure in aesthetic things?

Humans have been coveting art for millennia. Today, the global fine-art market is worth more than US\$50 billion annually. Scientists are finally getting a grip on the biology behind our passion for objects — beautiful or otherwise.

No one brain region is involved in art appreciation. Instead, complex and widely distributed neural activity characterizes the aesthetic experience. Much of this activity is in the brain's reward circuitry, which also responds to drugs, sex and attractive faces.

Researchers still do not agree on the definition of an aesthetic experience, much less on how the brain regions work together to create it. Answers will require coordination between many scientific fields, as well as art theory and philosophy.

Chelsea Ward is a freelance science writer in Vienna, Austria.

LETTERS TO THE EDITOR

Rewards of beauty: the opioid system mediates social motivation in humans

Molecular Psychiatry (2014) **19**, 746–747; doi:10.1038/mp.2014.1; published online 11 February 2014

Facial attractiveness is a powerful cue that affects social communication and motivates sexual behavior.^{1–3} Attractive people are both judged⁴ and treated⁵ more positively, reflecting the biased stereotypical notion that 'beautiful is good'. Indeed, beautiful faces are processed by the limbic reward system⁶ and according to the same economic principles as non-social rewards.⁷ The human reward system has a high density of μ -opioid receptors,⁸ which have an important role in affiliation and attachment.^{9–11} Here, we causally test whether the healthy human opioid system mediates facial attractiveness preference.

In rodents, μ -opioid (MOR) neurotransmission can increase both hedonic value ('liking') and motivational salience ('wanting') of rewards.¹² When several rewards are available, MOR agonism increases and antagonism decreases preference specifically for the

most valuable option. For instance, rats ate fewer palatable cookies but not less standard chow after MOR antagonism,¹³ while MOR stimulation enhanced sexual 'wanting' of only estrous, but not nonestrous, females.¹⁴ We predicted that antagonism of the human opioid system would decrease, while MOR agonism would increase 'liking' and 'wanting'⁶ specifically for the evolutionarily most valuable option, namely attractive opposite-sex faces.

In this double-blind, placebo-controlled cross-over study, 30 healthy males (see Figure 1 and Supplementary Information) viewed photographs of faces of varying attractiveness levels. In each session, participants received a μ -opioid receptor agonist (morphine 10 mg), a nonselective opioid receptor antagonist (naltrexone 50 mg) or placebo, and performed one 'liking' and one 'wanting' task (see Parsons *et al.*¹⁵). In the 'liking' task, participants viewed each face for 5 s before rating attractiveness on a visual analog scale with anchors 'very unattractive' to 'very attractive'. In a 'wanting' task of fixed duration (3.65 min), participants increased or decreased the preset viewing time of each face (5 s) by pressing

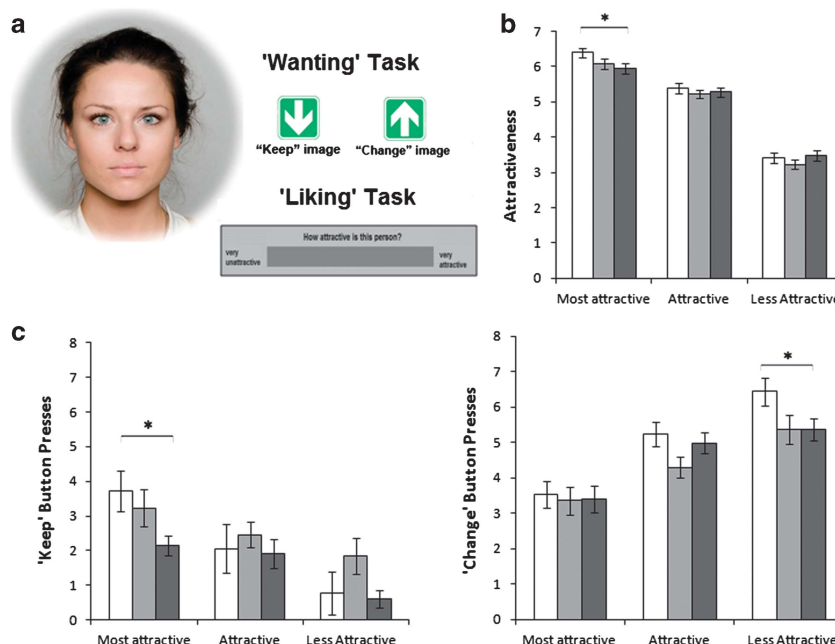


Figure 1. (a) In both the 'wanting' and 'liking' tasks, participants viewed faces from a purposely developed Oslo Face Database. The attractiveness categories were calculated based on the ratings from an independent group of 20 male participants (the depicted female face is from the 'Most Attractive' category). Thirty healthy males received a μ -opioid agonist (morphine 10 mg), a nonselective opioid antagonist (naltrexone 50 mg) or placebo (per-oral) on three separate days. In the 'wanting' task, participants could press one of two arrow keys to view the image for longer ('keep') or shorter ('change') time than the preset 5 s, without altering total task duration. To account for the data loss due to technical error of the 'liking' task, seven more participants were tested using the same paradigm. In the 'liking' task, participants rated attractiveness of each face using a VAS scale. (b) Morphine treatment enhanced and naltrexone treatment decreased men's 'liking' of the most beautiful female faces (ratings on the VAS scale from 0 to 10). (c) 'Wanting' of attractive females, as measured by the total of 'keep' button presses, was similarly affected by opioid manipulations. However, morphine also increased motivation to avoid viewing the least attractive female faces, as measured by the total of 'change' button presses. * $P < 0.05$.

buttons to keep looking at the same face or change by proceeding to the next face.¹⁵

Linear multilevel (mixed models) regression analysis was employed to assess viewing times and 2.5-SD-trimmed 'keep' or 'change' button-press data from the 'wanting' task, and attractiveness ratings from the 'liking' task. Main factors were drug, attractiveness and gaze direction. Control variables were session number, stimulus order, image set and OPRM1 group (AA or GA; see Supplementary Information).

'WANTING' TASK

Both morphine and naltrexone decreased the average viewing time by ~200 ms compared with placebo (main effect of drug, $F_{(2,1208)} = 3.6$, $P = 0.026$). However, a planned contrast of 'keep' button-press data revealed the expected pattern of 'wanting' increases with morphine and decreases with naltrexone for the most attractive female faces ($M > N$, $t = 2.56$, $P = 0.011$, Cohen's $d = 0.95$, Figure 1b). Yet, for the least attractive females, morphine increased 'wanting' to change the photo, relative to placebo and naltrexone treatment ($M > N$, $t = 2.52$, $P = 0.012$, Cohen's $d = 0.94$). Analysis of total key-presses per image revealed a significant increase in 'wanting' behavior after morphine relative to naltrexone treatment ($F_{(2,1206)} = 5.2$, $P = 0.006$, $M > N$, $t = 3.18$, $P = 0.001$, Cohen's $d = 1.18$), consistent with opioid mediation of human motivational preference for faces.

'LIKING' TASK

In line with our prediction that the opioid system mediates 'liking' for the evolutionarily most valuable option, attractiveness ratings were significantly higher after morphine compared with naltrexone treatment only for the most attractive female faces ($M > N$, $t = 2.13$, $P = 0.034$, Cohen's $d = 0.91$, Figure 1b, data from 23 participants, see Supplementary Information). The main effect of drug did not reach significance ($F_{(2,1314)} = 1.8$, $P = 0.16$).

Our results offer the first evidence that pharmacological manipulation of the human MOR system affects both aesthetic evaluation of and motivation for viewing opposite-sex faces. In line with findings from rodent literature,^{13,14} the effects of the MOR manipulations were strongest for the most valuable stimuli, that is, the most beautiful women. Morphine increased and naltrexone decreased men's 'liking' of these faces. We also observed an increase in 'wanting' behavior after morphine relative to naltrexone treatment, indicating that manipulation of the opioid system affected participants' motivation to expend effort. Specifically, activation of the opioid reward system with morphine not only increased 'wanting' key-press behavior to keep viewing the beautiful faces but also increased motivation to avoid viewing the least attractive faces.

The two components of reward, hedonic evaluation ('liking') and motivational salience ('wanting') were previously shown to partially dissociate when men viewed female faces of varying attractiveness levels,⁶ and when males and females viewed images of infants.¹⁵ The current study revealed an opioid-related increase in motivation to avoid the least attractive female faces, which was not mirrored by changes in reported attractiveness of these faces. For the most beautiful faces, however, the MOR manipulations affected 'liking' and 'wanting' similarly and in the expected directions.¹¹ Together, these findings suggest that the human opioid system may mediate social motivation by enhancing the salience and reward appraisal of the most valuable stimuli, while inhibiting 'wanting' of less valuable social cues.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

We thank MH Sneve, T Karlsson and S Aminihajibashi for technical assistance; I Olsen, L Bachs, V Vindenes and E Øiestad for pharmacological advice; J Gjerstad for help with genotyping; and Drs M Kringelbach, B Bastian, L Thomsen and G Overskeid for helpful comments on earlier drafts of this manuscript. The project was funded by grant number E5455867 to S Leknes from the Research Council of Norway.

O Chelnokova¹, B Laeng¹, M Eikemo¹, J Riegels¹, G Løseth¹, H Maurud¹, F Willoch² and S Leknes^{1,2}

¹Department of Psychology, University of Oslo, Oslo, Norway and

²Department of Medicine, University of Oslo, Oslo, Norway

E-mail: o.v.chelnokova@psykologi.uio.no or

s.g.leknes@psykologi.uio.no

REFERENCES

- Perrett DI, Lee KJ, Penton-Voak I, Rowland D, Yoshikawa S, Burt DM *et al.* *Nature* 1998; **394**: 884–887.
- Rhodes G, Simmons LW, Peters M. *Evol Hum Behav* 2005; **26**: 186–201.
- Parsons CE, Young KS, Mohseni H, Woolrich MW, Thomsen KR, Joensson M *et al.* *Soc Neurosci-Uk* 2013; **8**: 268–274.
- Dion K, Walster E, Berschei. E. *J Pers Soc Psychol* 1972; **24**: 285–289.
- Langlois JH, Kalakanis L, Rubenstein AJ, Larson A, Hallam M, Smoot M. *Psychol Bull* 2000; **126**: 390–423.
- Aharon I, Etcoff N, Ariely D, Chabris CF, O'Connor E, Breiter HC. *Neuron* 2001; **32**: 537–551.
- Hayden BY, Parikh PC, Deane RO, Platt ML. *Proc Biol Sci* 2007; **274**: 1751–1756.
- Biederman I, Vessel EA. *Am Sci* 2006; **94**: 247–253.
- Nelson EE, Panksepp J. *Neurosci Biobehav Rev* 1998; **22**: 437–452.
- Machin AJ, Dunbar RIM. *Behaviour* 2011; **148**: 985–1025.
- Hsu DT, Sanford BJ, Meyers KK, Love TM, Hazlett KE, Wang H *et al.* *Mol psychiatry* 2013; **18**: 1211–1217.
- Berridge KC, Kringelbach ML. *Psychopharmacology* 2008; **199**: 457–480.
- Cooper SJ, Turkish S. *Pharmacol Biochem Be* 1989; **33**: 17–20.
- Mahler SV, Berridge KC. *Psychopharmacology* 2012; **221**: 407–426.
- Parsons CE, Young KS, Kumari N, Stein A, Kringelbach ML. *Plos ONE* 2011; **6**: e20632.

Supplementary Information accompanies the paper on the Molecular Psychiatry website (<http://www.nature.com/mp>)

Metamoodics: meta-analysis and bioinformatics resource for mood disorders

Molecular Psychiatry (2014) **19**, 747–749; doi:10.1038/mp.2013.118; published online 10 September 2013

Mood disorders, including major depression (MD) and bipolar disorder (BP), are among the most common psychiatric disorders.¹ Although genetic factors have an important role in their etiology,^{2,3} the heritability of these disorders remains largely unexplained. Owing to advances in high-throughput genomic technologies, it is becoming increasingly feasible to investigate the genetic architecture of these disorders at an unprecedented resolution. The challenge is how to make sense of the enormous flood of data generated by the diverse genomic studies that use these new

Abnormal Brain Network Organization in Body Dysmorphic Disorder

Donatello Arienzo¹, Alex Leow^{1,2,3}, Jesse A Brown⁴, Liang Zhan⁵, Johnson GadElkarim^{1,6}, Sarit Hovav⁷ and Jamie D Feusner^{*7}

¹Department of Psychiatry, University of Illinois, Chicago, Chicago, IL, USA; ²Department of Bioengineering, University of Illinois, Chicago, Chicago, IL, USA; ³Community Psychiatry, Sacramento, CA, USA; ⁴Center for Cognitive Neuroscience, University of California, Los Angeles, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA; ⁵Imaging Genetics Center, Laboratory of Neuro Imaging, Department of Neurology, University of California, Los Angeles, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA; ⁶Department of Electrical and Computer Engineering, University of Illinois, Chicago, Chicago, IL, USA; ⁷Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA

Body dysmorphic disorder (BDD) is characterized by preoccupation with misperceived defects of appearance, causing significant distress and disability. Previous studies suggest abnormalities in information processing characterized by greater local relative to global processing. The purpose of this study was to probe whole-brain and regional white matter network organization in BDD, and to relate this to specific metrics of symptomatology. We acquired diffusion-weighted 34-direction MR images from 14 unmedicated participants with DSM-IV BDD and 16 healthy controls, from which we conducted whole-brain deterministic diffusion tensor imaging tractography. We then constructed white matter structural connectivity matrices to derive whole-brain and regional graph theory metrics, which we compared between groups. Within the BDD group, we additionally correlated these metrics with scores on psychometric measures of BDD symptom severity as well as poor insight/delusional. The BDD group showed higher whole-brain mean clustering coefficient than controls. Global efficiency negatively correlated with BDD symptom severity. The BDD group demonstrated greater edge betweenness centrality for connections between the anterior temporal lobe and the occipital cortex, and between bilateral occipital poles. This represents the first brain network analysis in BDD. Results suggest disturbances in whole brain structural topological organization in BDD, in addition to correlations between clinical symptoms and network organization. There is also evidence of abnormal connectivity between regions involved in lower-order visual processing and higher-order visual and emotional processing, as well as interhemispheric visual information transfer. These findings may relate to disturbances in information processing found in previous studies.

Neuropsychopharmacology (2013) **38**, 1130–1139; doi:10.1038/npp.2013.18; published online 13 February 2013

Keywords: graph theory; DTI; connectivity; clustering coefficient; global efficiency; betweenness centrality

INTRODUCTION

Body dysmorphic disorder (BDD) is characterized by preoccupation with misperceived defects of appearance or excessive concern about slight physical anomalies, causing clinically significant distress and impairment of functioning (American Psychiatric Association, 2000). BDD affects approximately 2% of the general population, making it

more prevalent than schizophrenia or bipolar I disorder (Buhlmann *et al*, 2010; Koran *et al*, 2008). It is associated with high lifetime rates of psychiatric hospitalization (48%) (Phillips and Diaz, 1997a) and suicide attempts (22.2–27.5%) (Buhlmann *et al*, 2010; Phillips *et al*, 2005). Insight in BDD is on a continuum, with 35.6–60% of BDD patients being delusional in their convictions of disfigurement (Mancuso *et al*, 2010; Phillips *et al*, 2006). Despite its high prevalence and severity, relatively little is known about the neurobiology.

Individuals with BDD perceive details of appearance features as defective without seemingly being able to contextualize that they are minor relative to their whole appearance. Moreover, neuropsychological (Deckersbach *et al*, 2000) and psychophysical (Feusner *et al*, 2010a; Stangier *et al*, 2008) studies suggest greater local relative to global visual and visuospatial processing. Functional magnetic resonance imaging (fMRI) studies using own and others' faces and inanimate object stimuli also suggest imbalances in detailed

*Correspondence: Dr JD Feusner, Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles, David Geffen School of Medicine at UCLA, 300 UCLA Medical Plaza, Suite 2200, Los Angeles, CA 90095, USA, Tel: +1 310 206 4951, Fax: +1 323 443 3593, E-mail: jfeusner@mednet.ucla.edu

Results presented in part at the American College of Neuropsychopharmacology 50th Annual Meeting, Waikoloa, HA, USA, December 2011

Received 12 October 2012; revised 26 December 2012; accepted 2 January 2013; accepted article preview online 15 January 2013

vs holistic/configural processing marked by abnormalities in primary and/or secondary visual cortical, temporal, and prefrontal systems (Feusner *et al*, 2007, 2010c, 2011).

Thus, there is evidence suggesting abnormalities in information processing in BDD. However, multiple interacting systems are likely responsible for neuropsychological and psychophysical performance (ie, visual perceptual systems, attentional systems, and prefrontal executive function systems); from these studies it is therefore difficult to discern which system(s) may be operating abnormally. In addition, the fMRI studies have reported regional abnormalities, rather than what is occurring on a network level, or how regions interact within larger systems to process information.

To better understand brain network organization in BDD, this study investigated structural networks using a graph-theoretical approach. This provides quantitative analyses of complex brain networks by modeling them as organizational systems, which can additionally be related to information such as clinical symptom severity. Structural connectivity patterns may predict functional connectivity patterns (Honey *et al*, 2009; Kotter and Sommer, 2000); thus, structural network topology may provide indirect information about functional organization in BDD. No study to date has investigated brain network organization in BDD.

Here, we seek to characterize whole brain and regional white matter network organization in individuals with BDD relative to that in healthy controls, and to relate this organization to clinical symptom severity. The phenomenology, as well as neuropsychological, psychophysical, and functional neuroimaging studies informed our hypotheses.

We hypothesized that the whole brain network organization in individuals with BDD would reflect highly localized information processing. This would manifest in: (1) highly localized subnetworks, with resultant abnormally high modularity (Fan *et al*, 2011); (2) less efficient transfer of information across the whole brain, resulting in lower global efficiency (Bullmore and Sporns, 2009); and (3) abnormally high mean clustering coefficient (MCC) (similar to what was previously found in a study in a related disorder, obsessive-compulsive disorder (OCD) (Zhang *et al*, 2011)). Additionally, we hypothesized that BDD symptom severity and poor insight/delusionalty would positively correlate with mean CC and modularity, and negatively correlate with global efficiency.

We also predicted abnormalities in the BDD group in regions involved in visual and emotional processing, and in frontostriatal systems. First, as informed by a previous fMRI study showing hypoactivity in dorsal visual stream regions (Feusner *et al*, 2007) (which contribute to holistic/configural visual processing), we predicted lower connectivity of nodes in the dorsal visual stream with other nodes in the brain. Specifically, the BDD group would have lower node degree in the superior parietal lobule, lateral occipital cortex, cuneus, supramarginal gyrus, and angular gyrus (Creem and Proffitt, 2001). Second, we hypothesized lower node degree in the left lingual gyrus, left occipital pole, and left occipital fusiform gyrus, which are regions found to be hypoactive in a previous study of own-face processing (Feusner *et al*, 2010c). Third, owing to the finding in that study of hyperactivity within frontostriatal circuits (orbito-

frontal cortex (OFC) and the caudate), we predicted higher node betweenness centrality. This would reflect greater influence of these frontostriatal regions with respect to the whole network; we postulated that this would be due to a dominant effect on the network of engagement of obsessive thoughts and compulsive behaviors, which were previously found to correlate with activity in these regions (Feusner *et al*, 2010c).

Finally, we hypothesized lower edge betweenness centrality for node pairs connecting regions in the anterior temporal lobe with regions in the occipital lobe, and node pairs connecting right and left occipital cortices. These hypotheses were informed by findings from a previous diffusion tensor imaging (DTI) study in which fiber disorganization in the inferior longitudinal fasciculus (ILF) and forceps major (FM) correlated with the clinical symptom of poor insight (Li *et al*, 2010). (For a more detailed description of these graph theory metrics, see Supplementary Materials and methods.)

MATERIALS AND METHODS

Participants

The UCLA Office of Human Research Protection Program approved the study protocol. In all, 14 unmedicated participants with BDD and 16 healthy controls, ages 20–48 years old, provided informed consent and were enrolled. BDD and control participants of equivalent gender, age, and level of education were recruited from the community. All had previously participated in a prior fMRI study of own-face processing (Feusner *et al*, 2010b). All were right-handed, as determined through the Edinburgh Handedness Inventory (Oldfield, 1971). Diagnoses were made by JDF, who has clinical expertise with this population, using the BDD Module (Phillips, 1995), a reliable diagnostic module modeled after the Structured Clinical Interview for DSM-IV. In addition, we performed a comprehensive clinical psychiatric evaluation and screened BDD and healthy control participants for comorbid Axis I disorders with the Mini-International Neuropsychiatric Interview (MINI) (Sheehan *et al*, 1998).

The following served as exclusion criteria: substance abuse and/or dependence within the past 12 months, lifetime neurological disorder, pregnancy, or any current medical disorder that may affect cerebral metabolism. We excluded BDD participants with any concurrent axis I disorder besides dysthymic disorder, major depressive disorder (MDD), or generalized anxiety disorder (GAD). Depression and anxiety are frequently comorbid in BDD, and thus a sample excluding these would not be representative. However, we required that BDD be the primary diagnosis as defined by the MINI. Healthy controls could not have any current or past axis I disorder. To assess BDD symptom severity, we administered the BDD version of the Yale-Brown Obsessive-Compulsive Scale (BDD-YBOCS) (Phillips *et al*, 1997b), a validated scale widely used to evaluate symptom severity in BDD, with scores ranging from 0 to 48. To assess insight and delusionalty, we administered the Brown Assessment of Beliefs Scale (BABS), a validated scale with scores ranging from 0 to 24 (Eisen *et al*, 1998). Higher BABS scores index poorer insight. Last,

we used the 17-item Hamilton Depression Rating Scale (HDRS) (Hamilton, 1960), and the Hamilton Anxiety Rating Scale (HARS) (Hamilton, 1969), both widely used and well-validated scales, to measure depressive and anxiety symptoms, respectively.

All BDD participants were required to have a score of ≥ 20 on the BDD-YBOCS, were free from psychotropic medications for a minimum of 8 weeks before study entry, and were not receiving cognitive-behavioral therapy.

Imaging Data Acquisition

We scanned participants using a 3 T Allegra MRI scanner (Siemens Medical Solutions USA Inc., Malvern, Pennsylvania). Diffusion-weighted MR imaging data using single-shot spin-echo echo-planar imaging were acquired using the following parameters: field of view = 240 mm; voxel size = $2.5 \times 2.5 \times 3.0 \text{ mm}^3$, with 0.75 mm gap; TR/TE = 7400/96 ms; and flip angle 9° . We collected 44 contiguous axial slices aligned

to the anterior commissure–posterior commissure line along 34 gradient-sensitizing directions with $b = 1000 \text{ s/mm}^2$ and one minimally diffusion-weighted scan. In addition, high-resolution structural images were acquired using T1-weighted magnetization-prepared rapid gradient echo (MP-RAGE) with the following parameters: sagittal slicing; TR = 2300 ms; TE = 293 ms; matrix = 256×256 ; 160 slices; 0.5 mm gap; field of view = $256 \times 256 \times 160 \text{ mm}^3$; flip angle = 8° ; and voxel size = $1.3 \times 1.3 \times 1.0 \text{ mm}^3$.

Data Processing

Figure 1 illustrates the data processing steps.

Calculation of diffusion tensors. All DTI data were corrected for eddy current and motion distortions using FSL (http://www.fmrib.ox.ac.uk/fsl/fdt/fdt_eddy.html), and the gradient table was updated based on the computed rotation matrix. We used DTIFIT in FSL

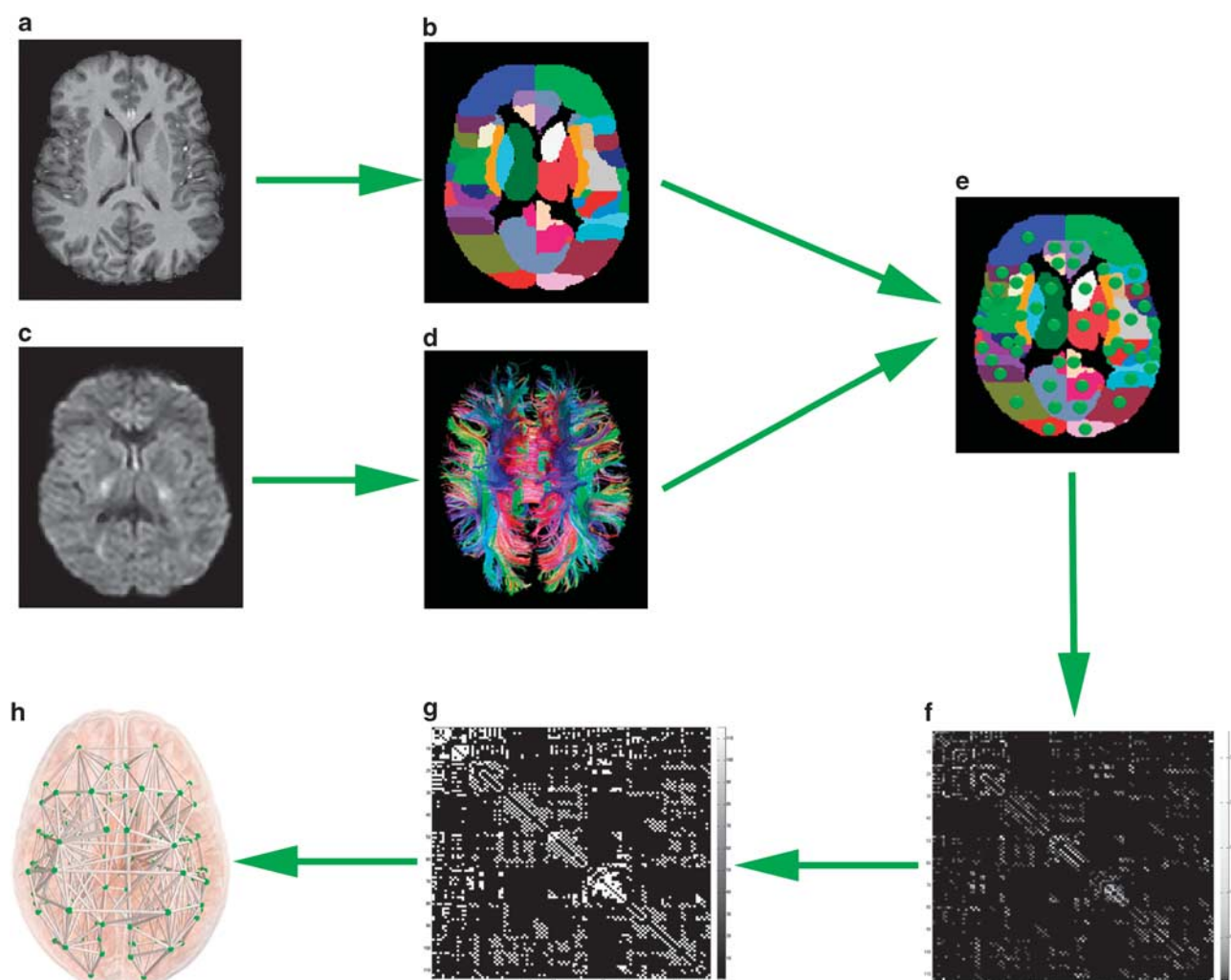


Figure 1 Structural network analysis processing. (a) High-resolution T1 magnetization-prepared rapid gradient echo (MP-RAGE) of an individual example participant. (b) In all, 113 cortical and subcortical regions of interest (ROIs), covering the whole brain, from the Harvard–Oxford probabilistic atlas. (c) Diffusion-weighted magnetic resonance imaging (MRI). (d) Whole brain tractography. (e) ROIs were registered to each participant's diffusion tensor imaging (DTI) space, and served as nodes from which the number of streamlines between them was identified. (f) A total of 113×113 weighted matrix and (g) 113×113 binarized matrix. (h) Network render of an individual participant imposed on the Montreal Neurological Institute (MNI) standard brain (for visualization purposes this only shows the top 6% strongest connections).

(http://www.fmrib.ox.ac.uk/fsl/fdt/fdt_dtfitt.html) to fit a diffusion tensor at each voxel.

DTI deterministic tractography. We computed whole-brain deterministic DTI tractography using Diffusion Toolkit (<http://trackvis.org/blog/tag/diffusion-toolkit/>). We reconstructed white matter fiber tracts by seeding at every voxel in the brain and applying the Fiber Assignment by Continuous Tracking (FACT) algorithm (Mori and van Zijl, 2002) with a maximum turn angle of 35°. Cortical and subcortical regions of interest (ROIs) were defined using the Harvard–Oxford cortical and subcortical probabilistic atlases (Desikan *et al*, 2006). All midline cortical masks were bisected to define separate hemispheric ROIs for each cortical region. The masks were set to a liberal probabilistic threshold of 10% to allow for the inclusion of tissue along the gray–white matter interface, where DTI tractography estimates are most reliable (Morgan *et al*, 2009). We used FSL’s FLIRT program (Jenkinson *et al*, 2002) to determine the optimal transformation between each participant’s DTI volume and the corresponding MP-RAGE (12 degree-of-freedom (d.f.) affine registration with a mutual information-based cost function), as well as between each participant’s MP-RAGE and the MNI152 T1 average brain (on which the Harvard–Oxford probabilistic atlases are based). We then combined the obtained two transformations to yield a final transformation, which was subsequently inverted and applied to register the 113 ROIs (in the atlas space) to each participant’s DTI space. To assure that ROI masks did not overlap after registration, each voxel was uniquely assigned to the mask for which it had the highest probability of membership.

Matrix construction. For each pair of ROIs, we determined the number of fibers connecting them. A fiber was considered to connect two ROIs if it originated in the first ROI and terminated in the second, or *vice versa*. We repeated this process for all possible pairs to determine the whole brain fiber connectivity matrix. These matrices served as the input for subsequent brain network analyses (Rubinov and Sporns, 2010). We assessed the matrices at 11 different sparsity levels (defined as the existing number of edges in a graph divided by the maximum possible number of edges) from 10 to 20%, at intervals of 1%. (For further description of, and rationale for, this sparsity thresholding, see Supplementary Materials and methods.) We then binarized the thresholded matrices to create corresponding brain network adjacency matrices, where 1 represents a connection and 0 represents no connection.

Graph Theory Metrics

The following is a brief description of the graph theory metrics used in this study. (For further descriptions, see Supplementary Materials and methods and Bullmore and Sporns (2009)) In graph theory, a network is comprised of ‘nodes’ (here, anatomically defined ROIs) and the connections or ‘edges’ between them (in this case, the white matter tracts). Node degree is the number of total nodes in the network that have direct connections to that node; a high value thus signifies that this node is highly connected to

other nodes in the network. The CC of a node is the ratio of the number of actual connections among its first-degree neighbors to the number of all possible connections. Thus, a high CC value for a node indicates that its neighbors are strongly interconnected to one another. MCC is the average of the CC for all nodes in the network; high values may confer greater local efficiency of information transfer of a network (Bullmore and Sporns, 2009). Globally efficiency is mathematically defined by averaging the inverse shortest path lengths across all node pairs. (Path length is computed by counting the minimum number of intermediate nodes needed to pass through to link any node pair.) A high global efficiency value represents a high overall capacity for parallel information transfer and integrated processing (Bullmore and Sporns, 2012). Modularity measures how strongly nodes in a community interconnect in comparison to a random graph. Thus, the higher the modularity value for a given community structure, the less likely it is to be the result of chance alone. Node betweenness centrality is the fraction of all shortest paths that contain a specific node. Thus, higher values indicate that a node has more ‘influence’ over flow of information between other nodes, in networks in which information tends to follow the shortest available path (Girvan and Newman, 2002). Similarly, edge betweenness centrality is the fraction of all shortest paths in the network that contain this connection. Higher values indicate a connection that has greater influence over other connections in the network.

We analyzed these connectivity matrices using the Brain Connectivity Toolbox (<https://sites.google.com/a/brain-connectivity-toolbox.net/bct/>) to yield the graph theory metrics of interest. For each metric we evaluated the area under the curve (AUC) over a range of sparsities to provide summarized measures of the network. We calculated both local metrics (for specific nodes) and global metrics (averaged across all nodes).

Three global network metrics were of primary interest based on the hypotheses of this study: CC, global efficiency, and modularity. Additionally, three local network metrics were of primary interest: node degree, node betweenness centrality, and edge betweenness centrality. For node degree we examined nodes in the dorsal visual stream (superior parietal, lateral occipital, cuneus, supramarginal gyrus, and angular gyrus) (Creem and Proffitt, 2001), as well as the left lingual gyrus, left occipital pole, and left occipital fusiform gyrus. For node betweenness centrality, we examined the OFC and caudate.

For edge betweenness centrality, we examined sets of connections between visual and emotional processing systems. These included node pairs between the anterior temporal lobe (temporal pole, amygdala, and hippocampus) and the occipital lobe (occipital fusiform, temporal occipital fusiform, and occipital pole), which approximates white matter connections via the ILF (Catani and Schotten, 2008). We also examined the node pair of the right and left occipital cortex (right and left occipital pole), which approximates connections via the FM (Catani and Schotten, 2008).

Statistical Analyses

We conducted statistical analyses on age- and gender-corrected data using General Linear Model Univariate in

SPSS, with gender as a fixed factor and age as continuous predictor.

For global network metrics, we performed two-tailed two-sample *t*-tests to compare MCC, global efficiency, and modularity AUC values between the healthy control and BDD groups. Because these metrics are non-independent, with respective separate hypotheses for each, we analyzed these separately rather than implementing an omnibus test or correcting for multiple comparisons. We used Pearson's correlation coefficients to assess the association between the global network metrics and BDD-YBOCS and BABS scores in the BDD group. We used a significance threshold of $\alpha = 0.05$, Bonferroni corrected for multiple comparisons.

For the local network metrics, we performed repeated-measures ANOVA to compare separately each graph theory metric of node degree, node betweenness centrality, and edge betweenness centrality values between the healthy control and BDD groups; group was one factor, and each node (or node pair for edge betweenness centrality) was the

repeated-measures factor. We used Huynh–Feldt adjustments for non-sphericity.

RESULTS

Demographics and Psychometrics

All BDD participants had preoccupations with perceived facial defects. Two had comorbid GAD, one had comorbid MDD, and three had both GAD and MDD or dysthymia (Table 1).

Global Network Results

The BDD group showed significantly higher MCC than controls across the range of sparsities (aside from at 17%) (Figure 2a), and for the AUC (5.14 ± 0.071 vs 5.05 ± 0.074 , $t = 3.5$, d.f. = 28, Cohen's $d = 1.32$, $P = 0.0015$) (Figure 2d). There were no significant differences between groups for any of the sparsity values or AUC for modularity (AUC: 4.3 ± 0.11 for BDD, 4.3 ± 0.12 for controls, $t = -0.12$,

Table 1 Demographics and Psychometric Scores

| Characteristic | BDD group (N = 14) | Control group (N = 16) | P-value ^a |
|------------------------------|--------------------|------------------------|----------------------|
| Age (years), mean (SD) | 26.7 (4.9) | 27.3 (5.3) | 0.75 |
| Female/male, no. | 7/7 | 8/8 | >0.99 |
| Education (years), mean (SD) | 15.5 (2.9) | 16.9 (2.3) | 0.150 |
| BDD-YBOCS score, mean (SD) | 29.6 (4.6) | N/A | N/A |
| BABS score, mean (SD) | 14.9 (4.1) | N/A | N/A |
| HDRS score, mean (SD) | 10.1 (6.7) | 1.3 (1.5) | <0.001 |
| HARS score, mean (SD) | 12.2 (7.7) | 1.6 (1.4) | <0.001 |

Abbreviations: BDD, body dysmorphic disorder; BDD-YBOCS, BDD version of the Yale–Brown Obsessive–Compulsive Scale; BABS, Brown Assessment of Beliefs Scale; HDRS, 17-item Hamilton Depression Rating Scale; HARS, Hamilton Anxiety Rating Scale; N/A, not applicable.

^aTwo-sample *t*-tests for age, education, and HDRS; χ^2 test for female/male.

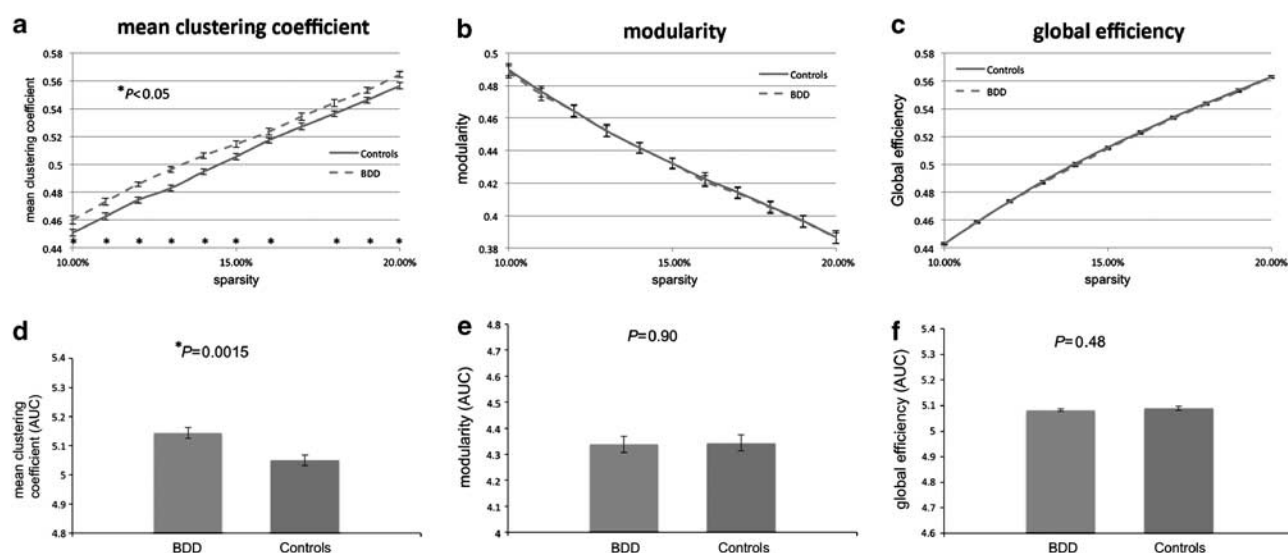


Figure 2 Global network measures in body dysmorphic disorder (BDD) and healthy control groups. Graphs in the top row show between-group differences as a function of network sparsity threshold for: (a) mean clustering coefficient (MCC); (b) modularity; and (c) global efficiency. Graphs in the bottom row show between-group differences for the area under the curve (AUC) for the range of sparsities tested for: (d) MCC; (e) modularity; and (f) global efficiency.

d.f. = 28, $P = 0.90$) (Figures 2b and e) or global efficiency (AUC: 5.08 ± 0.02 for BDD, 5.09 ± 0.03 for controls, $t = -0.71$, d.f. = 28, $P = 0.48$) (Figures 2c and f).

Correlation with Clinical Variables

There was a significant negative correlation between global efficiency and BDD-YBOCS scores ($r = -0.68$, $P = 0.0069$) (Figure 3). Modularity positively correlated with BABS scores ($r = 0.54$, $P = 0.047$), although it did not survive Bonferroni correction. There were no significant correlations between MCC and BDD-YBOCS ($r = 0.26$, $P = 0.37$) or BABS ($r = -0.074$, $P = 0.80$) scores, between global efficiency and BABS ($r = -0.24$, $P = 0.40$), or between modularity and BDD-YBOCS scores ($r = 0.15$, $P = 0.62$).

As a *post hoc* analysis, we explored separate correlations between the network metrics and the items of the BDD-YBOCS that index obsessional thoughts (items 1–5) and the items that index behaviors (compulsive-like and avoidant—items 6–10 and 12). There was a significant negative correlation between behaviors and global efficiency ($r = -0.70$, $P = 0.0047$), but the correlation between obsessive thoughts and global efficiency was not significant ($r = -0.48$, $P = 0.081$). There were no significant correlations with MCC and modularity for these subscale measures.

We also conducted *post hoc* correlation analyses to test the relationship between the network metrics and depression severity. (We did not test relationships with anxiety severity separately, as the correlation between HARS and HDRS scores in our sample was $r = 0.93$) There were no significant correlations between HDRS scores and modularity ($r = -0.46$, $P = 0.098$), global efficiency ($r = 0.036$, $P = 0.90$), or MCC ($r = 0.34$, $P = 0.23$).

Local (Nodal) Network Results

The analysis of edge betweenness centrality revealed a significant group effect ($F_{1,28} = 4.22$, $P = 0.049$) and node

effect ($F_{10.57,296} = 20.39$, $P < 0.0001$), but no group by node effect ($F_{10.57,296} = 1.15$, $P = 0.33$). *Post hoc* two-sample *t*-tests revealed that the BDD group showed significantly higher edge betweenness centrality for the connection between left temporal pole and left occipital pole (BDD-CON = 60.44, $t = 3.73$, d.f. = 28, Cohen's $d = 1.4$, $P = 0.00086$), the connections between left temporal pole and left temporal occipital fusiform cortex (BDD-CON = 23.76, $t = 2.14$, d.f. = 28, Cohen's $d = 0.8$, $P = 0.041$), between left amygdala and left occipital pole (BDD-CON = 30.15, $t = 2.2$, d.f. = 28, Cohen's $d = 0.83$, $P = 0.036$), and between right amygdala and right occipital pole (BDD-CON = 15.15, $t = 2.13$, d.f. = 28, Cohen's $d = 0.8$, $P = 0.042$). Only the connection between left temporal pole and left occipital pole survived Bonferroni correction, using a corrected α threshold of $0.05/18 = 0.0028$ (accounting for all 18 possible node connections for the right and left between the anterior temporal lobe and the visual cortex). The BDD group showed significantly higher edge betweenness centrality for the connection between left and right occipital pole (BDD-CON = 112.39, $t = 2.12$, d.f. = 28, Cohen's $d = 0.8$, $P = 0.043$).

There were no significant differences between groups for either node degree or node betweenness centrality (Supplementary Table S1).

We additionally conducted *post hoc* correlation analyses between edge betweenness centrality values and BDD-YBOCS scores and BABS scores for the connections that were significantly different between groups (before correction for multiple comparisons). For the connection between the left temporal pole and the left temporal occipital fusiform cortex, there was a significant correlation between edge betweenness centrality and total BDD-YBOCS scores ($r = 0.73$, $P = 0.0031$), which survived Bonferroni correction (α threshold of $0.05/10 = 0.005$). For this connection, there were significant correlations between both the BDD-YBOCS obsessive thoughts items and the behaviors items, and edge betweenness centrality ($r = 0.71$, $P = 0.0043$ and $r = 0.72$, $P = 0.0039$, respectively). There was also a correlation

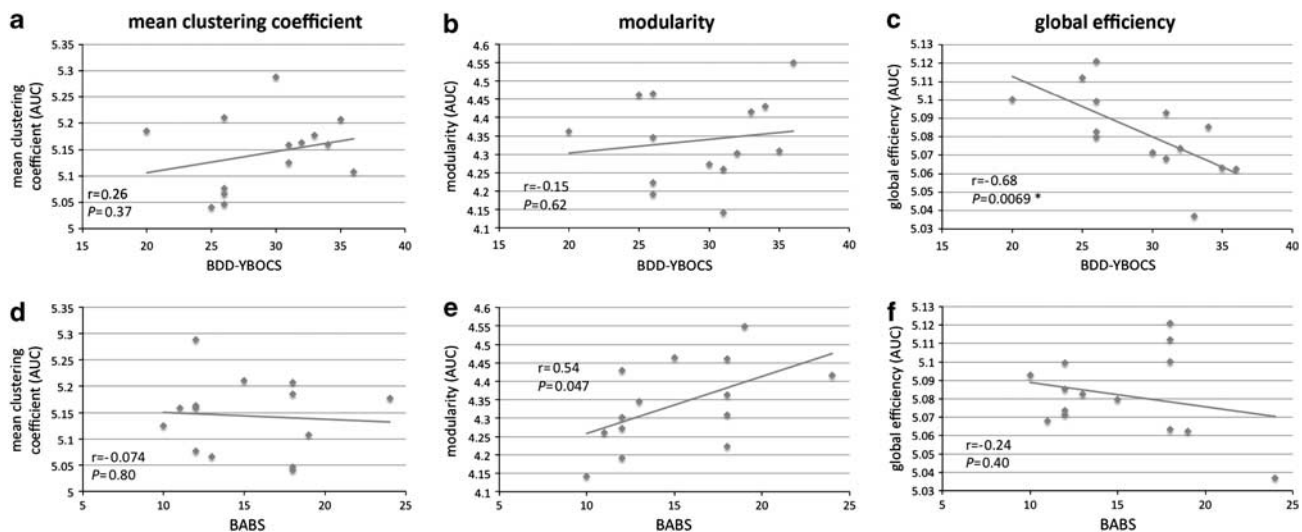


Figure 3 Correlations between clinical severity measures and global network measures in individuals with body dysmorphic disorder (BDD). The top row shows correlations between the BDD version of the Yale–Brown Obsessive Compulsive Scale (BDD-YBOCS) scores and area under the curve (AUC) values for: (a) mean clustering coefficient (MCC); (b) modularity; and (c) global efficiency. The bottom row shows correlations between the Brown Assessment of Beliefs Scale (BABS) scores (a measure of poor insight/delusional) and AUC values for: (d) MCC; (e) modularity; and (f) global efficiency.

between behaviors and edge betweenness centrality for the left temporal pole/left occipital fusiform cortex connection ($r=0.56$, $P=0.036$) (not surviving Bonferroni correction). The correlation between obsessive thoughts and edge betweenness centrality for this connection was not significant ($r=0.46$, $P=0.098$).

Additional Analyses

We additionally performed *post hoc* analyses to explore if prior treatment in BDD subjects affected results. Eight of the BDD subjects were treatment naïve, and six had received prior medication treatment. (Two subjects received brief psychotherapy, but it was unrelated to BDD.) The previously medicated group showed significantly higher MCC than controls (AUC: 5.19 ± 0.061 vs 5.05 ± 0.074 , $t=4.09$, d.f. = 20, Cohen's $d=1.83$, $P=0.00057$), while the BDD medication naïve group showed a trend for higher MCC than controls (AUC: 5.11 ± 0.059 vs 5.05 ± 0.074 , $t=3.5$, d.f. = 22, Cohen's $d=0.83$, $P=0.064$).

For the local (nodal) results, there were no significant differences between groups when we separately analyzed the BDD medication naïve and the previously medicated BDD groups, each compared with matched sets of healthy controls. (This may have been due to loss of power in these smaller subgroups.)

DISCUSSION

This study represents the first brain network analysis in BDD. Individuals with BDD exhibit abnormal white matter brain network organization, as characterized by higher MCC compared with controls. In addition, global efficiency negatively correlates with BDD symptom severity. Individuals with BDD also demonstrate higher edge betweenness centrality for connections between anterior temporal and occipital regions, as well as between bilateral occipital poles.

Global Metrics

As hypothesized, individuals with BDD have higher MCC relative to controls, suggesting a disturbance in network organization. In general, higher CCs are found in networks with a more regular, as opposed to random, organization (Stam and Reijneveld, 2007). Higher CCs are thought to confer locally higher degree of information transfer (Bullmore and Sporns, 2009). However, such networks that are overall more regular may also exhibit globally reduced signal propagation speed, computational power, and synchronizability across distant regions (Watts and Strogatz, 1998). As structural and functional network organizations in the brain share many topological features (Honey *et al*, 2010) (and structural connectivity patterns may predict functional connectivity patterns; Honey *et al* (2009), Honey *et al* (2010), and Kotter and Sommer (2000)), such disturbance in structural network organization may provide indirect information about functional organization in BDD.

Higher MCC in BDD suggests a network organization in which local connections dominate. One possible clinical implication of this could be an imbalance in global and local processing of visual information, leading to a distorted perception of appearance; individuals with BDD perceive

detailed imperfections and flaws (local information) and are unable to contextualize them as minor relative to their whole appearance. This is consistent with previous findings in BDD that provide evidence of greater local (relative to global) visual and visuospatial information processing (Deckersbach *et al*, 2000; Feusner *et al*, 2007, 2010a, c, 2011). Specifically, a previous neuropsychological study that included a measure of visuospatial construction and memory (Rey Osterrieth Complex Figure Test) demonstrated poor performance in the BDD group, mediated by poor organizational strategies due to selective recall of details instead of larger design features (Deckersbach *et al*, 2000). A study examining the face inversion effect found that individuals with BDD, relative to healthy controls, demonstrated less delay in response time when identifying upside-down relative to upright faces, suggesting a tendency to engage in highly detailed processing of faces regardless of their orientation. In contrast, healthy controls are more likely to engage in holistic processing of upright faces, yet rely on (slower) detailed processing of inverted faces. Another psychophysical experiment in BDD found an advantage for change detection for facial features of others' faces (Stangier *et al*, 2008).

However, it is unclear if abnormal performance in these studies was the result of impairments in information processing at the level of executive functioning resulting in poor memory organization strategies, selective attention, visual integration, or lower-order detail and/or holistic/configural visual processing. Higher MCC found in this study could reflect abnormal network organization across the whole brain, and could relate to these abnormalities in information processing; however, future studies are necessary to test this directly.

Although no previous brain network studies have been performed in BDD, one study examined functional network properties in OCD (Zhang *et al*, 2011). Relative to healthy controls, individuals with OCD demonstrated a pattern of significantly greater CCs, but not significantly different shortest path lengths. However, this was only found in a 'top-down control network' (multiple prefrontal, parietal, temporal, occipital, and subcortical regions) but not for whole-brain functional networks. The study also found that CC correlated with functional connectivity for primarily short-range functional connections. BDD has similarities to OCD in terms of overlapping phenomenology, shared heredity, and evidence of shared genetics (Hollander and Wong, 1995; Monzani *et al*, 2012). Although this study in OCD (Zhang *et al*, 2011) examined functional rather than structural networks, the current study suggests that BDD individuals may show similar patterns of aberrant structural network properties.

In this study, the previously medicated group had greater differences in MCC relative to healthy controls than the medication naïve group did. A possible explanation for this is that individuals who were more severely ill, and hence had more aberrant brain network organization, were more likely to have been medicated in the past.

Local Metrics

Contrary to our hypotheses, we found greater edge betweenness centrality for connections between temporal

pole and occipital pole nodes. White matter tracts connecting these regions are considered part of the ILF (Catani and Schotten, 2008). Significantly greater edge betweenness centrality was observed for nodes that span early (eg, V2) visual processing systems with higher-order visual processing systems in the temporal pole. The temporal pole is thought to be involved in integration of sensory, motor, and linguistic information with semantic knowledge and has been proposed to represent a hub in a cortical semantic network (Patterson *et al*, 2007). Because edge betweenness centrality quantifies the fraction of all shortest paths in a graph containing the given edge, greater values in BDD between occipital and anterior temporal regions suggests that this connection is more influential on the whole brain network in individuals with BDD compared with controls. This may also be indicative of heightened communication between these nodes.

A relevant clinical implication of this is that visual information processing for individuals with BDD may interfere or 'bleed into' many cognitive processes. This is consistent with the observation that for individuals with BDD a very large proportion of their time (on average 3–8 h per day (Phillips, 2005)) is occupied by intrusive, obsessive thoughts, usually relating to the visual perception of their appearance. The fact that the finding in this study was significant on the left may point to a more dominant left hemisphere involvement in visual information processing in BDD, consistent with existing evidence for greater detail and analytic processing found in a previous fMRI study (Feusner *et al*, 2007).

There are alternative interpretations of the findings involving the temporal pole, as it subserves multiple functions. For example, a recent study in pathological gambling found relationships between activity in the temporal poles and both gambling urges and subjective emotional responses (Balodis *et al*, 2012). Given previously described functions of the temporal pole, the authors offered the possible explanation that affectively salient gambling cues may have triggered the retrieval of personally relevant emotional memories. A similar process may occur in individuals with BDD, but in this case visual appearance cues may be the typical trigger of such emotional memories.

Contrary to our hypotheses, we found greater edge betweenness centrality for the connection between bilateral occipital poles. This connection likely approximates the portion of the FM that connects relatively early visual processing regions of the right and left visual cortical hemispheres (Putnam *et al*, 2010). Although the significance of this finding relative to the phenomenology of BDD is not entirely clear, a possible explanation is that preoccupations with perceived defects in BDD may rely heavily on visual processing, for which early communication between bilateral visual fields is generally important.

Correlations with Clinical Variables

In the BDD group, there was a significant negative correlation between global efficiency and BDD-YBOCS. Greater severity of symptoms is thus associated with lower global integration of the network. The results for the separate correlation analyses with the obsessive thoughts and the behaviors items of the BDD-YBOCS suggest that the

relationship between BDD symptom severity and global efficiency is driven to a greater extent by the severity of compulsive/avoidant behaviors than by obsessive thoughts. This suggests that BDD symptomatology (in particular compulsive and avoidant behaviors) globally relates to a multitude of brain subsystems. The overall effect of such interactions may be to impact negatively the efficiency of the nervous system.

Interestingly, despite this negative correlation, we did not detect a significant between-group difference in global efficiency. We posit that phenotypes, such as propensity for obsessive thoughts and compulsive behaviors, may map better to brain pathophysiology than DSM or ICD-10 diagnostic categories. Categorical diagnostic constructs in psychiatry, such as the diagnosis of BDD, may encompass multiple overlapping endophenotypes and phenotypes (Insel and Cuthbert, 2009), particularly as they may represent heterogeneous groupings of symptom clusters or dimensions.

Secondly, on a local level the strong correlation between edge betweenness centrality and BDD-YBOCS scores, for the connection between the left temporal pole and the left temporal occipital fusiform cortex, is an indication that individuals with greater severity of symptoms tend to have a larger percentage of all shortest paths that include this connection. As this connection is likely facilitated by the ILF, this strong correlation thus adds to a growing literature supporting the role of the ILF in feed-forward processes, which may involve consolidation of visual memories (Shinoura *et al*, 2007). Additionally, evidence also supports its involvement in feedback information processing, carrying signals regarding emotional valence of stimuli to the visual cortex and resulting in enhanced visual processing (Morris *et al*, 1998).

Moreover, evidence from non-human primates and humans suggests that the temporal pole is involved in linking highly processed perceptual information with emotional responses, which contributes to the formation of personal semantic memory (Olson *et al*, 2007). Thus, the degree to which those with BDD experience intrusive, obsessive thoughts and compulsive behaviors may be associated with the proportion of general information transfer throughout the brain that includes this connection, which is involved in integrating emotion and memory with visual processing systems.

A clinical implication of this is that in individuals with BDD, obsessive thoughts about appearance and, especially, compulsive behaviors and urges to engage in such behaviors, may be tightly linked with what they perceive visually. This, in turn, is influenced by visually related memories and emotion. Additionally, as this finding manifests with a statistically significant laterality (to the left side), a greater degree of detailed analytic visual processing may thus be associated with greater symptom severity (Evans *et al*, 2000). As such, in the future we plan to explore the possibility that this connection may represent an imaging biomarker for an important phenotype in BDD.

Limitations

Small sample size may have resulted in decreased ability to detect significant differences with smaller effect sizes. A

subset of the BDD group had comorbid GAD and/or a depressive disorder, which may have been a confound. Although we excluded substance abuse or dependence, we did not assess for tobacco use. Another limitation is that IQ measurements of participants were not available. Other studies have found IQ to be positively correlated with global efficiency in white matter networks (Li *et al*, 2009), and negatively correlated with path length (although not with MCC) in functional networks (van den Heuvel *et al*, 2009). In our study, the groups did not significantly differ on total years of education. Moreover, previous neuropsychological studies have not found abnormal IQ in individuals with BDD (Deckersbach *et al*, 2000; Dunai *et al*, 2010; Hanes, 1998).

In addition, there are inherent limitations in diffusion tractography (Jbabdi and Johansen-Berg, 2011). For example, such analyses may be dependent on the choice of anatomical atlas, which subsequently determine the choice the nodes (Wang *et al*, 2009; Zalesky *et al*, 2010). The 12-parameter affine transformations we used for realigning the MP-RAGE and DTI spaces only partially corrects for B0 inhomogeneity-induced geometric distortions; alternative techniques using nonlinear registration may better address such distortions for future studies.

CONCLUSIONS

Individuals with BDD show disturbances in topological organization of structural networks, which correlate with clinical symptomatology. In addition, there is evidence of abnormal connectivity in regions involved in interhemispheric visual information transfer, and those involved in lower- and higher-order visual and emotional processing, the latter of which also correlates with clinical symptomatology. These findings may be associated with disturbances in information processing found in previous studies. Future studies of individuals earlier in the course of illness (adolescence) and in unaffected first-degree relatives will be useful to determine if these findings represent endophenotypes predisposing to specific clinical symptoms in BDD.

ACKNOWLEDGEMENTS

This work was supported by grants from the National Institute of Mental Health (K23 MH079212 and R01MH093535, Dr Feusner).

DISCLOSURE

The authors declare no conflict of interest.

REFERENCES

- American Psychiatric Association (2000). *Diagnostic and Statistical Manual of Mental Disorders: DSM-IV-TR*. 4th edn. American Psychiatric Association: Washington, DC.
- Balodis IM, Lacadie CM, Potenza MN (2012). A preliminary study of the neural correlates of the intensities of self-reported gambling urges and emotions in men with pathological gambling. *J Gambl Stud* 28: 493–513.
- Buhlmann U, Glaesmer H, Mewes R, Fama JM, Wilhelm S, Brahler E *et al* (2010). Updates on the prevalence of body dysmorphic disorder: a population-based survey. *Psychiatr Res* 178: 171–175.
- Bullmore E, Sporns O (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat Rev Neurosci* 10: 186–198.
- Bullmore E, Sporns O (2012). The economy of brain network organization. *Nat Rev Neurosci* 13: 336–349.
- Catani M, Schotten Td (2008). A diffusion tensor imaging tractography atlas for virtual *in vivo* dissections. *Cortex* 44: 1105–1132.
- Creem SH, Proffitt DR (2001). Defining the cortical visual systems: 'What', 'Where', and 'How'. *Acta Psychol (Amst)* 107: 43–68.
- Deckersbach T, Savage C, Phillips K, Wilhelm S, Buhlmann U, Rauch S *et al* (2000). Characteristics of memory dysfunction in body dysmorphic disorder. *J Int Neuropsychol Soc* 6: 673–681.
- Desikan RS, Segonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D *et al* (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage* 31: 968–980.
- Dunai J, Labuschagne I, Castle DJ, Kyrios M, Rossell SL (2010). Executive function in body dysmorphic disorder. *Psychol Med* 40: 1541–1548.
- Eisen JL, Phillips KA, Baer L, Beer DA, Atala KD, Rasmussen SA (1998). The Brown Assessment of Beliefs Scale: reliability and validity. *Am J Psychiatry* 155: 102–108.
- Evans MA, Shedden JM, Hevenor SJ, Hahn MC (2000). The effect of variability of unattended information on global and local processing: evidence for lateralization at early stages of processing. *Neuropsychologia* 38: 225–239.
- Fan Y, Shi F, Smith JK, Lin W, Gilmore JH, Shen D (2011). Brain anatomical networks in early human brain development. *NeuroImage* 54: 1862–1871.
- Feusner JD, Hembacher E, Moller H, Moody TD (2011). Abnormalities of object visual processing in body dysmorphic disorder. *Psychol Med* 41: 2385–2397.
- Feusner JD, Moller H, Altstein L, Sugar C, Bookheimer S, Yoon J *et al* (2010a). Inverted face processing in body dysmorphic disorder. *J Psychiatr Res* 44: 1088–1094.
- Feusner JD, Moody T, Hembacher E, Townsend J, McKinley M, Moller H *et al* (2010b). Abnormalities of visual processing and frontostriatal systems in body dysmorphic disorder. *Arch Gen Psychiatry* 67: 197–205.
- Feusner JD, Moody T, Townsend J, McKinley M, Hembacher E, Moller H *et al* (2010c). Abnormalities of visual processing and frontostriatal systems in body dysmorphic disorder. *Arch Gen Psychiatry* 67: 197–205.
- Feusner JD, Townsend J, Bystritsky A, Bookheimer S (2007). Visual information processing of faces in body dysmorphic disorder. *Arch Gen Psychiatry* 64: 1417–1425.
- Girvan M, Newman ME (2002). Community structure in social and biological networks. *Proc Natl Acad Sci USA* 99: 7821–7826.
- Hamilton M (1969). Diagnosis and rating of anxiety. *Br J Psychiatry* 3: 76–79.
- Hamilton M (1960). A rating scale for depression. *J Neurol Neurosurg Psychiatry* 23: 56–62.
- Hanes K (1998). Neuropsychological performance in body dysmorphic disorder. *J Int Neuropsychol Soc* 4: 167–171.
- Hollander E, Wong C (1995). Introduction: obsessive-compulsive spectrum disorders. *J Clin Psychiatry* 56(Suppl 4): 3–6.
- Honey CJ, Sporns O, Cammoun L, Gigandet X, Thiran JP, Meuli R *et al* (2009). Predicting human resting-state functional connectivity from structural connectivity. *Proc Natl Acad Sci* 106: 2035–2040.
- Honey CJ, Thivierge J-P, Sporns O (2010). Can structure predict function in the human brain? *NeuroImage* 52: 766–776.
- Insel TR, Cuthbert BN (2009). Endophenotypes: bridging genomic complexity and disorder heterogeneity. *Biol Psychiatry* 66: 988–989.

- Jbabdi S, Johansen-Berg H (2011). Tractography: where do we go from here? *Brain Connect* 1: 169–183.
- Jenkinson M, Bannister P, Brady M, Smith S (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage* 17: 825–841.
- Koran LM, Abujaoude E, Large MD, Serpe RT (2008). The prevalence of body dysmorphic disorder in the United States adult population. *CNS Spectr* 13: 316–322.
- Kotter R, Sommer FT (2000). Global relationship between anatomical connectivity and activity propagation in the cerebral cortex. *Philos Trans R Soc Lond Ser B* 355: 127–134.
- Li W, Zhang L, Arienzo D, Leow A, Feusner JD (2010). Fractional anisotropy differences of the inferior fronto-occipital fasciculus in body dysmorphic disorder. *Society for Neuroscience Annual Meeting* San Diego, CA.
- Li Y, Liu Y, Li J, Qin W, Li K, Yu C et al (2009). Brain anatomical network and intelligence. *PLoS Comput Biol* 5: e1000395.
- Mancuso SG, Knoesen NP, Castle DJ (2010). Delusional versus nondelusional body dysmorphic disorder. *Compr Psychiatry* 51: 177–182.
- Monzani B, Rijdsdijk F, Iervolino AC, Anson M, Cherkas L, Mataix-Cols D (2012). Evidence for a genetic overlap between body dysmorphic concerns and obsessive-compulsive symptoms in an adult female community twin sample. *Am J Med Genet B* 159B: 376–382.
- Morgan VL, Mishra A, Newton AT, Gore JC, Ding Z (2009). Integrating functional and diffusion magnetic resonance imaging for analysis of structure-function relationship in the human language network. *PLoS One* 4: e6660.
- Mori S, van Zijl PC (2002). Fiber tracking: principles and strategies—a technical review. *NMR Biomed* 15: 468–480.
- Morris JS, Friston KJ, Buchel C, Frith CD, Young AW, Calder AJ et al (1998). A neuromodulatory role for the human amygdala in processing emotional facial expressions. *Brain* 121(Part 1): 47–57.
- Oldfield RC (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia* 9: 97–113.
- Olson IR, Plotzker A, Ezzyat Y (2007). The Enigmatic temporal pole: a review of findings on social and emotional processing. *Brain* 130: 1718–1731.
- Patterson K, Nestor PJ, Rogers TT (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nat Rev Neurosci* 8: 976–987.
- Phillips KA (2005). *The Broken Mirror Edition*. Oxford University Press: New York.
- Phillips KA (1995). Diagnostic Instruments for body dysmorphic disorder. *American Psychiatric Association 148th Annual Meeting* Miami, FL, p 157.
- Phillips KA, Coles ME, Menard W, Yen S, Fay C, Weisberg RB (2005). Suicidal ideation and suicide attempts in body dysmorphic disorder. *J Clin Psychiatry* 66: 717–725.
- Phillips KA, Diaz SF (1997a). Gender differences in body dysmorphic disorder. *J Nerv Ment Disord* 185: 570–577.
- Phillips KA, Hollander E, Rasmussen SA, Aronowitz BR, DeCaria C, Goodman WK (1997b). A severity rating scale for body dysmorphic disorder: development, reliability, and validity of a modified version of the Yale–Brown Obsessive Compulsive Scale. *Psychopharmacol Bull* 33: 17–22.
- Phillips KA, Menard W, Pagano ME, Fay C, Stout RL (2006). Delusional versus nondelusional body dysmorphic disorder: clinical features and course of illness. *J Psychiatr Res* 40: 95–104.
- Putnam MC, Steven MS, Doron KW, Riggall AC, Gazzaniga MS (2010). Cortical projection topography of the human splenium: hemispheric asymmetry and individual differences. *J Cogn Neurosci* 22: 1662–1669.
- Rubinov M, Sporns O (2010). Complex network measures of brain connectivity: uses and interpretations. *NeuroImage* 52: 1059–1069.
- Sheehan DV, Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E et al (1998). The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *J Clin Psychiatry* 59(Suppl 20): 22–33, quiz 34–57.
- Shinoura N, Suzuki Y, Tsukada M, Katsuki S, Yamada R, Tabei Y et al (2007). Impairment of inferior longitudinal fasciculus plays a role in visual memory disturbance. *Neurocase* 13: 127–130.
- Stam CJ, Reijneveld JC (2007). Graph theoretical analysis of complex networks in the brain. *Nonlinear Biomed Phys* 1: 3.
- Stangier U, Adam-Schwebe S, Muller T, Wolter M (2008). Discrimination of facial appearance stimuli in body dysmorphic disorder. *J Abnorm Psychol* 117: 435–443.
- van den Heuvel MP, Stam CJ, Kahn RS, Hulshoff Pol HE (2009). Efficiency of functional brain networks and intellectual performance. *J Neurosci* 29: 7619–7624.
- Wang J, Wang L, Zang Y, Yang H, Tang H, Gong Q et al (2009). Parcellation-dependent small-world brain functional networks: a resting-state fMRI study. *Hum Brain Mapp* 30: 1511–1523.
- Watts DJ, Strogatz SH (1998). Collective dynamics of ‘small-world’ networks. *Nature* 393: 440–442.
- Zalesky A, Fornito A, Harding IH, Cocchi L, Yucel M, Pantelis C et al (2010). Whole-brain anatomical networks: does the choice of nodes matter? *NeuroImage* 50: 970–983.
- Zhang TJ, Wang JH, Yang YC, Wu QZ, Li B, Chen L et al (2011). Abnormal small-world architecture of top-down control networks in obsessive-compulsive disorder. *J Psychiatr Neurosci* 36: 23–31.

Supplementary Information accompanies the paper on the Neuropsychopharmacology website (<http://www.nature.com/npp>)

ORIGINAL ARTICLE

Visual exposure to obesity: Experimental effects on attraction toward overweight men and mate choice in females

E Robinson and P Christiansen

BACKGROUND: Cultural differences in ideal body weight are well established, but less research has examined attraction toward potential mates of heavier body weights. We examined whether exposure to obesity increases physical attraction toward overweight men.

METHODS: In Studies 1 and 2, we examined the effect that exposure to obese vs healthy weight men had on female attraction toward an overweight man. Study 3 examined whether females who are regularly exposed to males of heavier body weights reported a greater attraction toward overweight men. Study 4 tested whether females in an online dating study were more likely to choose to date an overweight man, after having been exposed to obesity.

RESULTS: Exposure to obesity altered visual perceptions of what normal and therefore healthy body weights were and this resulted in greater attraction toward an overweight man (Studies 1 and 2). Females regularly exposed to men of heavier body weight reported a greater attraction toward overweight men (Study 3). After exposure to obesity, females in an online dating study were more likely to choose to date an overweight man ahead of a healthy weight man (Study 4).

CONCLUSIONS: Exposure to male obesity increases female attraction toward overweight men and may affect mate choice.

International Journal of Obesity (2015) 39, 1390–1394; doi:10.1038/ijo.2015.87

INTRODUCTION

Although men and women tend to find extremely high and low body mass indexes (BMI) less attractive than those closer to the 'normal' weight range,^{1,2} research has shown there are cross-cultural differences in preferences toward different body weights.³ There is some suggestion that social groups differ in their attitudes toward body weight, as ethnic groups with higher rates of adiposity have been shown to be less likely to view slender bodies as desirable.^{4,5}

One potential explanation of some of these cross-cultural differences in body size preferences may relate to the size of people an individual encounters in their social environment, whereby attitudes toward different body weights are flexible and adjust as a function of the body sizes a person is frequently exposed to. This proposition is supported by findings which indicate that visual perception can be recalibrated based on the types of stimuli a person is exposed to, commonly referred to as 'visual adaptation'.^{6–8}

Over the last 30 years, there has been a rapid increase in adiposity, with obesity now becoming common in many areas of the western world.^{9,10} Given that studies to date have tended to show that obese individuals are judged as undesirable partners^{11,12} and often presented in a negative way in popular media,^{13,14} one potential consequence could be that adiposity has become more unattractive and undesirable over time. However, a different hypothesis based on the visual adaptation literature^{15,16} is that because some individuals will be more frequently exposed to heavier body weights, this could make overweight individuals appear more appealing and attractive.

Recent work has suggested that as well as altering perceived normality of body weight,^{16,17} exposure to heavier body weights may recalibrate perceptions of what a 'healthy' body weight looks like.^{17,18}

Given that perceived health is an important cue to attraction and mate choice,^{19,20} it could be the case that exposure to heavier body weights makes overweight individuals appear more attractive, owing to frequent exposure altering perceived normality and healthiness of weight. Thus, if an individual was frequently being exposed to obesity in their social environment, this could have the result of making overweight individuals appear as more suitable or desirable potential mates.

To date, the effect that body weight exposure has on attraction toward overweight individuals or how exposure may alter mate choice has not been empirically examined. We examined these questions across four studies, all of which focused on female attraction toward overweight men. Study 1 tested whether exposure to obese, as opposed to healthy weight men, would result in females being more attracted to an overweight man. Study 2 examined whether this effect may be explained by a pathway involving exposure altering visual perceptions of both perceived normality and healthiness of weight. In Study 3, we tested the hypothesis that women who regularly socialise with men of heavier body weight (increased exposure to larger male body weights) would find overweight men more desirable. Finally, in Study 4, we invited single women to take part in an online dating study and tested whether exposing participants to images of obese males would later result in them being more likely to choose to date an overweight man, instead of a healthy weight man.

STUDY 1

Participants

White US female participants were recruited online from Amazon Mechanical Turk, in exchange for a small monetary reward.

Amazon Mechanical Turk is a validated crowdsourcing website used to recruit participants to take part in online research studies.²¹ The study was described as being about 'perceptions of other people'. We aimed to recruit a large sample size of approximately 500 participants for this initial study. Five hundred and thirteen participants took part (mean age = 32.3, s.d. = 10.8) and all participants completed the study. The sample had a mean BMI of 26.8 (s.d. = 7.7). BMI was calculated using self-reported weight and height in all studies. All studies reported were approved by the authors' institutional ethics board.

Stimuli

Participants were exposed to 10 photographs of obese men (obesity exposure condition), healthy weight men (healthy weight exposure condition) or everyday objects (control condition), before making judgements about an overweight man. Full length standardised photographs of Caucasian men aged 18–30 were used. Models were wearing normal fitting short-sleeved shirts and trousers or jeans, standing facing front with their arms at their sides, next to a standard sized door frame. No photographs of men who regularly participated in strength building sports or appeared to have muscular builds were used. Prior to Study 1, a pilot study ($n = 50$) was conducted to select a set of photographs of healthy weight men and a set of obese men (according to WHO BMI guidelines) who were matched for height, attractiveness, how muscular they appeared and how tight fitting their clothes were (see Robinson and Kirkham¹⁷ for more information about the photographs). In the present studies, we controlled for any interference of facial expression during exposure by obscuring the middle section of the models' faces with a black box. Facial features were not obscured for the overweight man, as we reasoned that rating attractiveness with facial features obscured may result in an artificial rating. The healthy weight models' mean BMI = 21 (range: 19.38–22.40) and the obese models' mean = 32 (range: 30.49–34.32). The overweight model's BMI = 27.

Procedure

After accessing the online site, participants were shown an information sheet and gave informed consent. Participants were then randomly assigned to the obesity, healthy weight or control exposure condition. To distract from the main aims of the study participants completed a set of mood items (for example, 'I am happy'). After this, participants were shown 10 photographs individually on separate pages and answered two questions about each image, using 7-point Likert scales. Participants in the obesity and healthy weight exposure conditions answered questions about the appearance of the men, although questions did not mention body weight or attraction (for example, 'this person looks relaxed'). The control condition answered questions about 10 images of everyday objects (for example, a sofa) using similar ratings as those made about the men (for example, 'this looks relaxing'). In all conditions, the eleventh photo was always of the overweight man. Underneath this photo, participants rated 'this person is attractive' and 'this man's clothes are bright' (to distract from our interest in body weight), using the same 7-point Likert scales. Participants then recorded demographic information, guessed the study aims and were debriefed.

Results

No participants guessed the aims of the study. SPSS 22 was used for analysis. Analyses showed that the conditions were balanced for age and BMI ($p > 0.05$). One-way analysis of variance indicated a significant effect of condition on how attracted to the overweight man participants were ($F^2 = 4.76$, $P = 0.009$, $f = 0.14$). Planned pairwise comparisons showed that participants exposed to obese men rated the overweight man as being significantly

more attractive than participants exposed to healthy weight men ($t(336) = 3.1$, $P = 0.002$, $d = 0.33$); see Figure 1. Compared with the object condition, participants exposed to the obese men rated the overweight man as being more attractive, although this difference was not statistically significant ($t(334) = 1.8$, $P = 0.079$, $d = 0.19$). Participants exposed to healthy weight men tended to rate the overweight man as being less attractive than the control condition, but again this difference was not significant ($t(340) = 1.3$, $P = 0.18$, $d = 0.15$). Thus, exposure to obese, as opposed to healthy weight men, influenced physical attractive toward an overweight man.

STUDY 2

Recent findings indicate that exposure to obesity alters normative perceptions of weight, which in turn also affects visual perceptions of what constitutes a healthy body weight.^{16,17} Health is an important cue that attraction is based upon.^{19,20} Thus, we hypothesised that the effect exposure to obesity has on attraction may be explained by exposure changing perceptions of weight normality and, in doing so, adjusting the extent to which an overweight man's weight appears healthy. We tested this in Study 2 using structural equation modelling.

Participants and procedure

UK women ($n = 137$) were recruited via social network sites (mean age = 22.8 years, s.d. = 3.5). This and the remaining studies were powered to detect a medium sized effect. The sample had a mean BMI of 23.0 (s.d. = 3.9). The exact same procedure was used as in Study 1, although instead of rating whether the overweight man was wearing a bright top, participants rated 'this person is a normal weight' (weight norm) and 'this person is a healthy weight' (perceived healthiness of weight). The order of the three questions was counterbalanced. We also opted to drop the object exposure condition, as our primary aim was to compare the effects of exposing participants to obese vs healthy weight men.

Results

No participants guessed the aims of the study. The two conditions were balanced for age and BMI ($P > 0.05$). *T*-tests indicated that after exposure to obesity, participants were more physically attracted to the overweight man, believed he was a more normal weight and a healthier weight, than after exposure to healthy weight men ($P < 0.001$). See Table 1 for full results. As hypothesised, the extent to which participants believed the man was a normal weight was correlated with perceptions of how healthy his weight was ($r = 0.72$, $P < 0.001$) and perceived

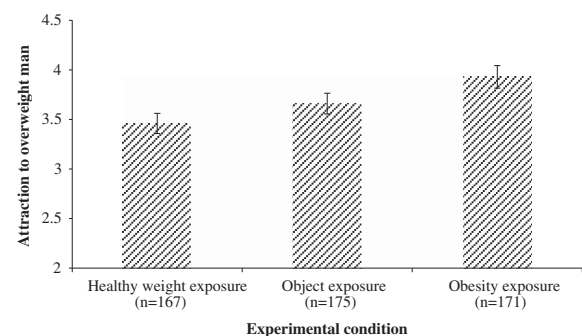


Figure 1. Attraction toward the overweight man by experimental condition, Study 1. Values are means. Error bars are standard error. Attraction to overweight man = 1–7 Likert scale ranging from strongly disagree to strongly agree. See text for statistical significance of between condition differences.

Table 1. Ratings of attraction, normality and healthiness of weight for the target overweight man in Study 2

| | Obesity exposure condition (n = 69) | Healthy weight exposure condition (n = 68) |
|-----------------------|-------------------------------------|--|
| Physical attraction | 3.8 (1.4) ^a | 3.0 (1.1) |
| Normality of weight | 5.7 (0.7) ^a | 4.3 (1.3) |
| Healthiness of weight | 5.7 (0.9) ^a | 4.1 (1.3) |

Values are means, with standard deviations in brackets. Measures = 1–7 Likert scale ranging from strongly disagree to strongly agree. Higher scores on measure denote greater attraction, perceived normality and healthiness of overweight man. ^aIndicates significant difference ($P < 0.001$) between two experimental conditions for measure.

healthiness of weight was also correlated with attraction ($r = 0.39$, $P < 0.001$), indicating that our proposed pathway may explain why exposure influences attraction.

We tested whether exposure altered attractiveness indirectly via the effect it had on judgements of weight normality and healthiness using structural equation modelling. When examining this pathway, we controlled for any effect that exposure may have on healthiness ratings directly, as well as examining the direct association between condition and attractiveness. We hypothesised that the exposure-attraction effect would be mediated by changes to normality of weight and then perceived healthiness of weight. We predicted that if this pathway was responsible for the exposure-attraction effect, then the relationship between exposure condition and attraction would become non-significant when accounting for this indirect pathway.

We confirmed model fit using a normed χ^2 ; whereby values between 1 and 3 are indicative of a good fit. The standardised root mean residual absolute fit index was also calculated, as well as the comparative fit index CFI. We also conducted a Bollen-Stine bootstrap as a final estimate of model fit. These indices indicated good model fit (see Figure 2). There was a significant indirect effect of condition on healthy weight via normal weight ($\beta = 0.30$; bootstrap estimates CI_{95} (confidence intervals): 0.05 to 0.12, $P = 0.002$), and a significant effect of normal weight on attraction via healthy weight ($\beta = 0.17$; bootstrap estimates CI_{95} : 0.07 to 0.48, $P = 0.003$). In line with our hypotheses, condition had a significant indirect effect on attraction via its effects on both normal weight and healthy weight ($\beta = 0.17$; bootstrap estimates CI_{95} : 0.02 to 0.12, $P = 0.003$). Moreover, the direct effect that condition had on attraction was no longer significant when accounting for the indirect pathway. These findings suggest that exposure to obesity may impact on attraction toward overweight men by causing adjustments to what is perceived as being a normal and therefore healthy body weight.

STUDY 3

Given our reliance on experimental designs in Studies 1 and 2, Study 3 examined whether cross-sectional data also support the notion that exposure to heavier male body weights increases attraction toward overweight men. We hypothesised that women who regularly socialise with heavier/overweight men (more frequent visual exposure) should prefer mates of heavier body weight.

Participants and procedure

Eighty UK undergraduate students (women) participated in a brief questionnaire study on a voluntary basis (mean age = 19.9 years, s.d. = 3.4), with a mean BMI of 21.4 (s.d. = 3.6). Ten men also took part in the study, but as all other studies we report examine female-male attraction, here we report analyses from only the

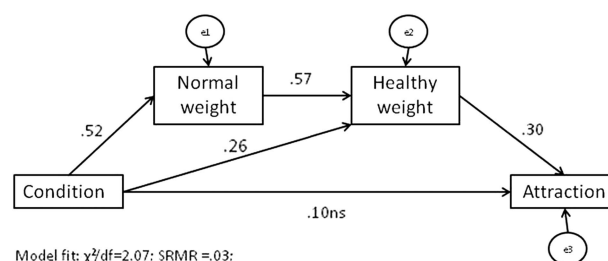


Figure 2. Study 2 pathway model. Figures are standardised regression coefficients and are significant at $P < 0.001$, unless otherwise stated (ns).

women. The inclusion of the 10 men does not change the results reported.

As a cover story, participants were asked to complete a series of questionnaires about attitudes to body weight and demographic information; including gender, age and self-reported weight and height (see Oldham and Robinson²²). Participants were shown nine body silhouette drawings of men, which ranged from very thin to very fat and increased incrementally.²³ Participants were asked to select which of the nine they thought was most attractive for a man (body weight attraction measure). On a different page, participants were shown an overweight man (as in Study 1) and rated 'compared to most other young males I spend time with, this person is' on a 10 cm visual analogue scale with anchors: far lighter, far heavier (size of male peers).

Results

To assess the association between participants' usual exposure to heavier body weights (size of male peers) and which of the nine body weight silhouettes they found most attractive (ordinal data), we used a Spearman's correlation coefficient. A significant association between the size of participants' peers and body weight attraction measure was observed ($r = 0.29$, $P = 0.01$), whereby spending time with heavier male peers was associated with participants selecting a larger body silhouette as being most attractive for a man.

Study 3 supports the experimental findings of Studies 1 and 2 by showing that women who report socialising with heavy men (more frequent exposure to heavier body weights) chose a heavier body weight as being most attractive for a man. Although this finding is in line with our hypotheses, the inverse may explain some of this association; women who find heavier bodies more attractive adopt heavy male friends. The aim of our final study was to test whether exposure could alter mate choice. To examine this, we tested whether exposure to obese men would cause participants to favour dating an overweight man over a healthy weight man, in an online dating study.

STUDY 4

The study was advertised as an online dating study. In studies 1–3, we did not use inclusion criteria for relationship status. Because our outcome measure of interest in Study 3 was mate choice, we recruited only *single* white females to participate who were of a similar age to the men presented in the study, to make dating choices as realistic as possible. This was specified in the study advert, along with a request for 18–30 year old participants.

Participants

One hundred and twenty-five UK single white women with mean age = 24.7 years (s.d. = 4.3) and mean BMI = 24.4 (s.d. = 5.7) were recruited online in exchange for entry into a small prize draw.

Procedure

After accessing the online site, participants were shown the 10 photographs of either obese or healthy weight men (Study 1 and 2). In keeping with the online dating study story, during the exposure phase, participants were asked to rate whether they would consider dating or be romantically interested in each of the men. After this initial exposure phase, participants were shown a photograph of an overweight man (Study 1) and next to this a photograph of a healthy weight man (BMI = 21.5) of similar height. Participants were asked to select which of the two men they would prefer to date. The position of the two photographs (left or right) was counterbalanced.

Results

No participants guessed the study aims. Conditions were balanced for age and BMI ($P > 0.05$). A chi-square test determined whether the exposure condition participants were assigned to influenced dating choice. Participants in the obesity exposure condition were more likely to choose to date the overweight man over the healthy weight man than participants in the healthy weight exposure condition ($\chi^2 = 4.2$, $P = 0.04$, $\Phi = 0.18$). Participants in the obesity exposure condition ($n = 64$) showed a preference toward dating the overweight man, choosing him 68.8% of the time, whereas only 50.8% of participants in the healthy weight exposure condition ($n = 61$) selected the overweight man. Thus, exposure to obese men resulted in participants being more likely to choose to date an overweight man, ahead of a healthy weight man.

THE EFFECT OF PARTICIPANT BMI

We also conducted additional analyses across all four studies to examine whether participant BMI (calculated using self-reported data) may be an important factor shaping attraction to overweight men. Across the studies, heavier women generally tended to find overweight men more attractive. The effect that exposing participants to obese men had on attraction was not moderated by participant BMI in any of the studies. See online Supplementary material for analyses. However, it is of importance to note that these patterns of results may have differed if we had included an objective measurement of participant body weight or a measurement of perceived personal (participant) weight status.

GENERAL DISCUSSION

Across four studies, we examined whether exposure to obese men increased female attraction toward overweight men. In Study 1, visual exposure to obese men resulted in women finding an overweight man more physically attractive. Study 2 showed that exposure to obese men altered visual perceptions of what normal and therefore healthy body weights were and this pathway mediated the effect that exposure had on attraction toward an overweight man. Study 3 conceptually replicated these findings by showing that women who regularly socialise with heavier men report a greater attraction toward overweight men. Finally, Study 4 showed that exposure to obese men, as opposed to healthy weight men, may have an effect on mate choice. Single women taking part in an online dating study were more likely to report that they would choose to date an overweight man ahead of a healthy weight man, after having been exposed to obese men.

The present studies show for the first time that visual exposure to heavier male body weights can shape how attractive females find men of heavier body weights. These findings are in line with suggestions that body weight preferences are likely to be determined by learning and environmental input.²⁴ In our studies, we speculate that exposure to obese men skewed the degree to which participants perceived the overweight model's weight as being 'normal', resulting in the overweight man in our studies

probably being perceived as being closer to what constitutes the mid-point of a distribution of male body weight.^{25,26}

The present findings also have relevance to societal increases in body weight. Previous work has shown that obese individuals are commonly seen as less desirable mates,^{11,12} but based on the present findings, one might predict that mate preferences concerning overweight individuals may become less negative. Here, we examined *visual* attraction toward *overweight* men, which is more common and presumably more socially acceptable than obesity, so it may be the case that exposure does not enhance attraction toward obese men. Given that obesity is now becoming prevalent in many parts of the world, we may start to see shifts in mate preference toward partners of slightly heavier body weights. The potential role that stigma has in explaining the present findings warrants attention. Study 2 showed that the exposure-attraction effect may be explained by exposure changing how healthy an overweight man's weight appeared. Although we interpret this in terms of health being used as a marker for mate suitability, it could also be the case that exposure resulted in participants being less likely to identify and label the man as being 'overweight'.²² This may have removed stigma associated with this label²⁷ and increased attractiveness.

The context in which participants were exposed to obesity in the present study also warrants discussion. Research indicates that obesity stigma may have increased in recent years,²⁸ and one consequence of this is that obese individuals are often portrayed in a negative manner in the media,¹³ as opposed to the neutral manner of exposure used in the present studies. Thus, it may be the case that the type of visual exposure to obesity which occurs through viewing popular media does not normalise heavier body weights.

Gender

The present studies were designed to examine the effect that exposure to different body weights has on *female* attraction to *overweight men*. One of our main reasons to do so was because we reasoned that there are likely to be less clearly defined standards for male body weight, in comparison with the widely internalised 'thin ideal' for female body weight.^{29,30} Thus, physical attraction toward overweight men, as opposed to overweight women, may be more sensitive to visual learning in an experimental context. As we observed consistent evidence that exposure to obese men was associated with increased female attraction toward an overweight man, this now raises a question of whether a similar pattern of results would be observed when examining *male* attraction toward *overweight women*. There is some research showing that attitudes toward female figures can be adjusted as a result of repeated exposure to slim or rounder figures,^{15,16} and women's views about their own body weight can be affected by visual exposure to thin vs fat females in the media.³¹ Thus, it seems conceivable that the body weight exposure effects observed in the present studies may translate to judgements made about women.

Strengths and limitations

The use of standardized images of actual healthy, overweight and obese individuals is a strength of this work, as often body image studies rely on artificial or perceptually extreme images,^{15,16} making it difficult to ascertain whether findings will transfer over to judgements made about actual people. Given that we examined attraction and mate choice shortly after exposure to a relatively small number of bodies, it is not clear from the present studies how long these effects would last for. However, if daily experience were to continually 'top up' exposure, we presume these effects could in theory be long lasting. A final limitation was that we only examined attraction toward Caucasian men, so further research examining how exposure to heavier body weights

influence attraction amongst other ethnicities would now be interesting.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

ER and PC conceived the experiments, analysed data and were involved in writing the paper. ER conducted the experiments. This research received no external funding. ER was partly supported by the Wellcome Trust.

REFERENCES

- Tov   M, Reinhardt S, Emery JL, Cornelissen PL. Optimum body-mass index and sexual attractiveness. *Lancet* 1998; **352**: 548.
- Tov   MJ, Cornelissen PL. Female and male perceptions of female physical attractiveness in front-view and profile. *Br J Psychol* 2001; **92**: 391–402.
- Marlowe F, Apicella C, Reed D. Men's preferences for women's profile waist-to-hip ratio in two societies. *Evol Hum Behav* 2005; **26**: 458–468.
- Rucker CE, Cash TF. Body images, body-size perceptions, and eating behaviors among African-American and white college women. *Int J Eat Disord* 1992; **12**: 291–299.
- Glasser CL, Robnett B, Feliciano C. Internet daters' body type preferences: race-ethnic and gender differences. *Sex Roles* 2009; **61**: 14–33.
- Webster MA, MacLeod DIA. Visual adaptation and face perception. *Philos Trans R Soc Lond B Biol Sci* 2011; **366**: 1702–1725.
- Rhodes G, Jeffery L, Watson T, Clifford CWG, Nakayama K. Fitting the mind to the world: Face adaptation and attractiveness aftereffects. *Psychol Sci* 2003; **14**: 558–566.
- Winkler C, Rhodes G. Perceptual adaptation affects attractiveness of female bodies. *Br J Psychol* 2005; **96**: 141–154.
- Ogden CL, Carroll M, Curtin LR, McDowell MA, Tabak CJ, Flegal KM. Prevalence of overweight and obesity in the United States, 199–2004. *J Am Med Assoc* 2006; **295**: 1549–1555.
- Burke MA, Heiland FW, Nadler CM. From overweight to about right: evidence of a generational shift in body weight norms. *Obesity (Silver Spring)* 2009; **18**: 1226–1234.
- Chen EY, Brown M. Obesity stigma in sexual relationships. *Obes Res* 2005; **13**: 1393–1397.
- Sitton S, Blanchard S. Men's preferences in romantic partners: obesity vs addiction. *Psychol Rep* 1995; **77**: 1185–1186.
- Puhl RM, Heuer CA. The stigma of obesity: a review and update. *Obesity (Silver Spring)* 2009; **17**: 941–964.
- Greenberg BS, Eastin ME, Hofschire L, Lachlan K, Brownell KD. Portrayals of overweight and obese individuals on commercial television. *Am J Pub Health* 2003; **93**: 1342–1348.
- Mele S, Cazzato V, Urgesi C. The importance of perceptual experience in the esthetic appreciation of the body. *PLoS One* 2013; **8**: e81378.
- Boothroyd LG, Tovee MJ, Pollet TV. Visual diet versus associative learning as mechanisms of change in body size preferences. *PLoS One* 2012; **7**: e48691.
- Robinson E, Kirkham TC. Is he a healthy weight? Exposure to obesity changes perceptions of what a healthy weight looks like. *Int J Obes (Lond)* 2014; **38**: 663–667.
- Robinson E, Christiansen P. The changing face of obesity: exposure to and acceptance of obesity. *Obesity (Silver Spring)* 2014; **22**: 1380–1386.
- Weeden J, Sabin J. Physical attractiveness and health in western societies: a review. *Psychol Bull* 2005; **131**: 635–653.
- Jones BC, Little AC, Burt DM, Perrett DI. When facial attractiveness is only skin deep. *Perception* 2004; **33**: 569–576.
- Berinsky AJ, Huber GA, Lenz GS. Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis* 2012; **20**: 351–368.
- Oldham M, Robinson E. Visual weight status misperceptions of men: Why overweight can look like a healthy weight. *J Health Psychol* e-pub ahead of print 20 January 2015.
- Stunkard AJ, Sorenson T, Schlusinger F. Use of the Danish adoption register for the study of obesity and thinness. In Kety S (ed). *The Genetics of Neurological and Psychiatric Disorders*. Raven Press: New York, 1980; 115–120.
- Tovee MJ, Edmonds L, Vuong QC. Categorical perception of human female physical attractiveness and health. *Evol Hum Behav* 2012; **33**: 85–93.
- Wedell DH, Santoyo EM, Pettibone JC. The thick and the thin of it: Contextual effects in body perception. *Basic Appl Soc Psych* 2005; **27**: 213–227.
- Melrose KL, Brown GDA, Wood AM. Am I abnormal? Relative rank and social norm effects in judgments of anxiety and depression symptom severity. *J Behav Decis Making* 2003; **26**: 174–184.
- Brochu PM, Esses VM. What's in a name? The effects of the labels 'fat' versus 'overweight' on weight bias. *J Appl Soc Psychol* 2011; **41**: 1981–2008.
- Andreyeva T, Puhl RM, Brownell KD. Changes in perceived weight discrimination among Americans: 1995–1996 through 2004–2006. *Obesity (Silver Spring)* 2008; **16**: 1129–1134.
- Rodin J. Cultural and Psychosocial determinants of weight concerns. *Ann Intern Med* 1993; **119**: 643–645.
- Hargreaves DA, Tiggemann M. Females 'thin ideal' media images and boys' attitudes towards girls. *Sex Roles* 2003; **49**: 539–544.
- Glauert R, Rhodes G, Byrne S, Fink B, Grammer K. Body size dissatisfaction and the effects of perceptual exposure on body norms and ideals. *Int J Eat Disord* 2009; **42**: 443–452.

Supplementary Information accompanies this paper on International Journal of Obesity website (<http://www.nature.com/ijo>)

ARTICLE

Received 17 Apr 2014 | Accepted 24 Jul 2014 | Published 16 Sep 2014

DOI: 10.1038/ncomms5800

Morphological and population genomic evidence that human faces have evolved to signal individual identity

Michael J. Sheehan¹ & Michael W. Nachman¹

Facial recognition plays a key role in human interactions, and there has been great interest in understanding the evolution of human abilities for individual recognition and tracking social relationships. Individual recognition requires sufficient cognitive abilities and phenotypic diversity within a population for discrimination to be possible. Despite the importance of facial recognition in humans, the evolution of facial identity has received little attention. Here we demonstrate that faces evolved to signal individual identity under negative frequency-dependent selection. Faces show elevated phenotypic variation and lower between-trait correlations compared with other traits. Regions surrounding face-associated single nucleotide polymorphisms show elevated diversity consistent with frequency-dependent selection. Genetic variation maintained by identity signalling tends to be shared across populations and, for some loci, predates the origin of *Homo sapiens*. Studies of human social evolution tend to emphasize cognitive adaptations, but we show that social evolution has shaped patterns of human phenotypic and genetic diversity as well.

¹Museum of Comparative Zoology and Integrative Biology, University of California, 3101 Valley Life Science Building, Berkeley, California 94720, USA. Correspondence and requests for materials should be addressed to M.J.S. (email: msheehan@berkeley.edu).

Human societies are predicated on our abilities to individually recognize and track scores of people in our social networks^{1,2}. The complexity of human societies is widely recognized as a major selective force that has shaped our cognitive abilities and social intelligence^{3–5}. Indeed, humans have highly developed individual recognition abilities, and there is a rich literature examining social cognition and individual recognition in humans^{6–9}. In particular, facial recognition plays a critical role in human social interactions, and the cognitive mechanisms underlying facial recognition have been well studied^{9–11}. When it comes to recognition, however, cognition is only one half of the equation. Recognition also depends on phenotypic variation in a population¹², without which discrimination is impossible. Compared with other animals or other parts of our bodies, we perceive human faces as being unusually variable and easy to recognize (Fig. 1). While this phenomenon can be at least partly explained by our specialization for learning human faces¹⁰, the fact that facial recognition is so important for social interactions among humans suggests that selection may have lead to increased facial distinctiveness. Despite the striking differences among human faces and the importance of facial identity for human society, the evolution of individuality in human faces has yet to be explored.

Theoretical and empirical studies have proposed multiple non-mutually exclusive negative frequency-dependent selection (NFDS) pressures that may maintain elevated phenotypic diversity in natural populations. These include apostatic selection, in which predation is lower on rare prey^{13,14}, mating preferences for novel phenotypes¹⁵ and selection to be recognizable¹⁶. For example, both apostatic selection and frequency-dependent mate preferences have been shown to contribute to the maintenance of the highly variable and heritable male colouration patterns in the guppy (*Poecilia reticulata*)^{14,17}. Frequency-dependent attractiveness has been proposed as a mechanism to explain patterns of hair and eye colour diversity in Europeans¹⁸ and recent tests have found empirical support for frequency-dependent attractiveness of beards in human males¹⁹. Although elevated variation in guppy colouration is limited to adult males¹⁷, human facial individuality is not limited to a particular age or sex class, suggesting that frequency-dependent mate preferences are unlikely to be the sole or major driver of elevated diversity in human facial patterns. Indeed, individual recognition is important in humans from cradle to grave across multiple

contexts. Selection to be individually recognizable in a variety of scenarios is therefore a prime hypothesis to explain the high diversity in human facial appearance¹².

Individual recognition will only evolve when it is beneficial to identify other individuals¹². Whether or not individuals benefit by being identifiable and easily recognized raises a different question. Traits used for individual recognition are expected to evolve as either identity cues or as identity signals depending on the benefits of being recognized^{12,20,21}. Identity cues are traits that allow discrimination but have not evolved for the purpose of recognition²² and are not expected to show signatures of adaptive evolution. Cues are essentially inadvertent phenotypic variation that other individuals can use for discrimination^{23,24}. For example, today human fingerprints are used for forensic identification though they have not evolved to facilitate recognition. As is the case for fingerprints, identity cues do not necessarily benefit individuals that are being recognized and may in fact harm them. In contrast, identity signals are traits that have been selected to facilitate individual recognition and as a result show elevated variation within populations^{16,25}. Individual recognition can rely on cues alone, but if individuals benefit on average from being recognized then selection is expected to favour individuals to advertise their identity with distinctive phenotypes^{12,20,21}. Identity signals evolve when being confused with others is costly due to misdirected behaviours including aggression²⁶, mating opportunities²⁷, parental care²⁸ and so on. Comparative and experimental evidence for identity signalling leading to increased phenotypic diversity has been documented in multiple taxa^{25,26,28,29} but has not been investigated in humans.

Individual recognition is facilitated when individuals display divergent trait values and novel combinations of traits^{11,21} leading to disruptive selection on multiple traits and the evolution of independent developmental pathways^{21,26}. Three key predictions of the identity-signalling hypothesis are (i) facial characteristics should be more variable than other visible traits not used for recognition, (ii) face traits are expected to show lower inter-trait correlations compared with other morphological traits and (iii) loci underlying normal facial variation are expected to show elevated genetic diversity consistent with NFDS favouring rare phenotypes. Loci contributing to identity-signalling traits are expected to show evidence of NFDS, such as an excess of intermediate-frequency alleles and elevated diversity when controlling for divergence^{30,31}. Selection for

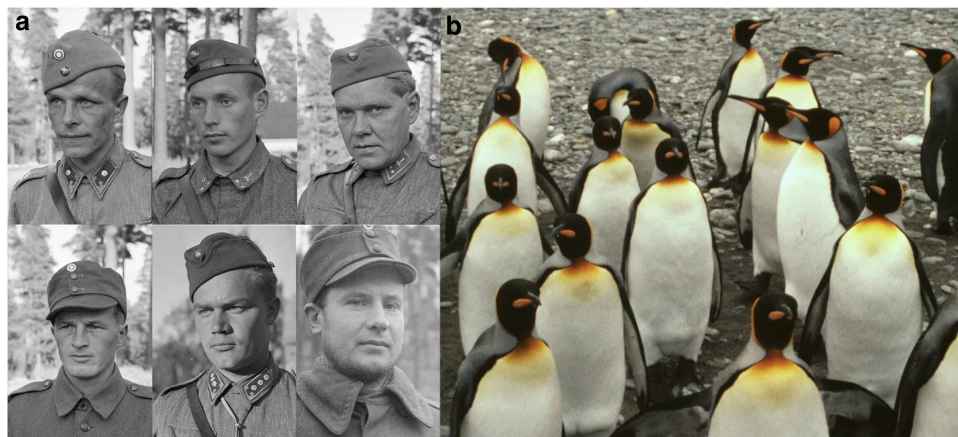


Figure 1 | Humans have much more individually distinctive faces than many animals. (a) Human populations show extensive variability in facial morphology that is used for individual recognition. Patterns of elevated variability are even maintained in more genetically homogeneous populations such as the Finnish, as demonstrated by the portraits of six male soldiers. **(b)** In contrast to the variability present in human faces, many animals such as king penguins have much more uniform appearances. While king penguins are not known to visually recognize individuals, they do have highly distinctive vocalizations that are used for individual recognition. (Photo credits: SA-kuva, Finnish Armed Forces photograph; Wikimedia commons.)

identity signalling on any one facial trait is likely to be relatively weak as there are numerous traits that contribute to individual identity. Due to the complex genetic architecture of facial variation^{32–34}, we expect a modest signature of elevated diversity in genomic regions underlying identity signals as a whole^{35,36}, though there may be stronger evidence for NFDS at a subset of loci. While it is plausible that identity cues could also show elevated phenotypic variation and reduced phenotypic correlations as a result of relaxed selection or stochastic developmental processes, the loci underlying identity cues are expected to evolve neutrally. Thus even a weak signature of NFDS, as may be expected for a complex quantitative trait such as facial identity, would reject the cue hypothesis and provide support for identity signalling.

Consistent with the predictions of the identity-signalling hypothesis, we find elevated phenotypic variation and reduced levels of inter-trait correlations in human faces compared with non-facial morphology. Furthermore, we find population genomic support for the identity-signalling hypothesis. Loci associated with variation in normal facial morphology show elevated nucleotide diversity compared with loci associated with variation in height or presumably neutral, intergenic variation. The loci with the strongest evidence of selection tend to be shared across continents, suggesting that selection on at least some loci associated with identity signalling is likely to be old. Indeed, by comparing sequences of modern humans with those of Neanderthals and Denisovans, we demonstrate that variation at some loci associated with facial morphology predates the origin of the human species. While studies of human social evolution have tended to emphasize its effect on cognition, our results suggest that social evolution has also played an important role in shaping human morphology.

Results

Morphological evidence. Morphological comparisons between faces and other traits are consistent with the predictions of identity signalling. We tested these predictions using data for 18 facial and 46 non-facial linear distance measures from the ANSUR anthropometric study of US army personnel³⁷ for females and males of African and European American ancestry, respectively (Supplementary Tables 1 and 2). Linear distances between facial landmarks have higher coefficients of variation than linear measurements of body traits in every group (Fig. 2a, Mann–Whitney *U* (MWU)-test, $n = 18$ facial and 46 non-facial measures, $P < 0.03$ for all comparisons). Without selection for uncoupled development, traits within individuals are generally correlated, as larger individuals tend to have larger traits^{38,39}. However, facial traits show lower inter-trait correlation coefficients than body traits in all four groups as predicted by the identity-signalling hypothesis (Fig. 2b, MWU-test, $n = 153$ facial correlations and 1,035 non-facial correlations, $P < 0.001$ for all comparisons). Indeed, the vast majority of the body measures are correlated (percentage of significant Pearson's correlations between traits, African American females (AAF) = 95.17%, African American males (AAM) = 99.14%, European American females (EAF) = 96.62%, European American males (EAM) = 99.81%, $n = 1,035$ pairwise comparisons), though many fewer facial traits are correlated with each other (AAF = 63.4%, AAM = 73.9%, EAF = 47.1%, EAM = 84.2%, $n = 153$ pairwise comparisons, Z -ratio < -11 and $P < 0.002$ for all comparisons). Uncorrelated values for face traits increase the diversity of facial phenotypes, facilitating recognition. For example, the breadth and length of hands are correlated (Fig. 2c, $r^2 = 0.30$, $P < 0.0001$), though the breadth and length of noses are not (Fig. 2d, $r^2 = 0.002$, $P = 0.06$). These results add

to previous findings that humans have among the lowest craniofacial morphological integration among primates and mammals more broadly⁴⁰.

Population genomic evidence. Using data from the 1000 Genomes Project⁴¹, we tested for a signature of NFDS in genomic regions surrounding single nucleotide polymorphisms (SNPs) previously associated with differences in normal facial morphology in Europeans^{32,33}. We compared the distribution of population genetic summary statistics calculated around face SNPs with the distribution of summary statistics for 5,000 putatively neutral intergenic regions as identified by the Neutral Region Explorer⁴². Here we present an analysis based on 2-kb windows, though the patterns of diversity reported here are robust to a range of window sizes (Supplementary Fig. 1). Elevated diversity in face regions relative to the intergenic regions would be consistent with the predictions of NFDS. However, it is possible that morphological traits in general could show elevated diversity in comparison with neutral regions, so we also compared face regions with regions surrounding SNPs associated with height⁴³, another complex morphological trait, as an additional control.

Patterns of diversity surrounding face-associated SNPs are consistent with NFDS on complex quantitative traits as predicted under the identity-signalling hypothesis. Here we present the values for the Finnish population (Fig. 3), though broadly similar overall patterns are found for other 1000 Genomes population samples from Europe and Africa and to a weaker extent Asia (Supplementary Figs 2–9). The folded site frequency spectrum shows that regions surrounding face-associated SNPs have an excess of intermediate-frequency variants compared with the two sets of control loci (Fig. 3a, MWU, $P < 0.0001$ for both comparisons). Additionally, the distribution of summary statistics for faces differs from distributions found for height or intergenic regions consistent with NFDS on facial traits (Fig. 3b–e, $P < 0.05$ for all comparisons). One possible confounding factor is that intermediate-frequency SNPs are over-represented in genotyping panels⁴⁴ and thus more often associated with traits in genome-wide association studies, so elevated diversity could conceivably be confounded by ascertainment biases. Two lines of evidence argue against this. First, the association studies and population genomic analyses were conducted in different samples of Europeans, and the patterns of elevated diversity around face SNPs are also found in African and Asian populations. Second, in the Finnish examined here, the minor allele frequency of the focal SNPs is actually lower for faces than for height (MWU, $P = 0.028$, Supplementary Fig. 10), suggesting that potential biases in association studies cannot explain the elevated patterns of diversity surrounding face-associated SNPs. The combination of elevated morphological and genetic diversity associated with human faces rejects a neutral explanation for human facial individuality and instead supports the hypothesis that human facial diversity is the product of selection for identity signalling in humans.

Evolutionary dynamics of identity-signalling loci. Due to the lack of data on SNPs associated with signalling traits in animals, population genomic methods have not previously been used to empirically explore the evolutionary dynamics of signalling traits. The present data set on identity signals in humans, however, provides an unprecedented opportunity to examine the history of selection on signalling traits used in social communication. Selection for identity signalling is expected to act on faces in all populations though it need not occur at the same loci. Conceivably, selection may act on the same loci across populations;

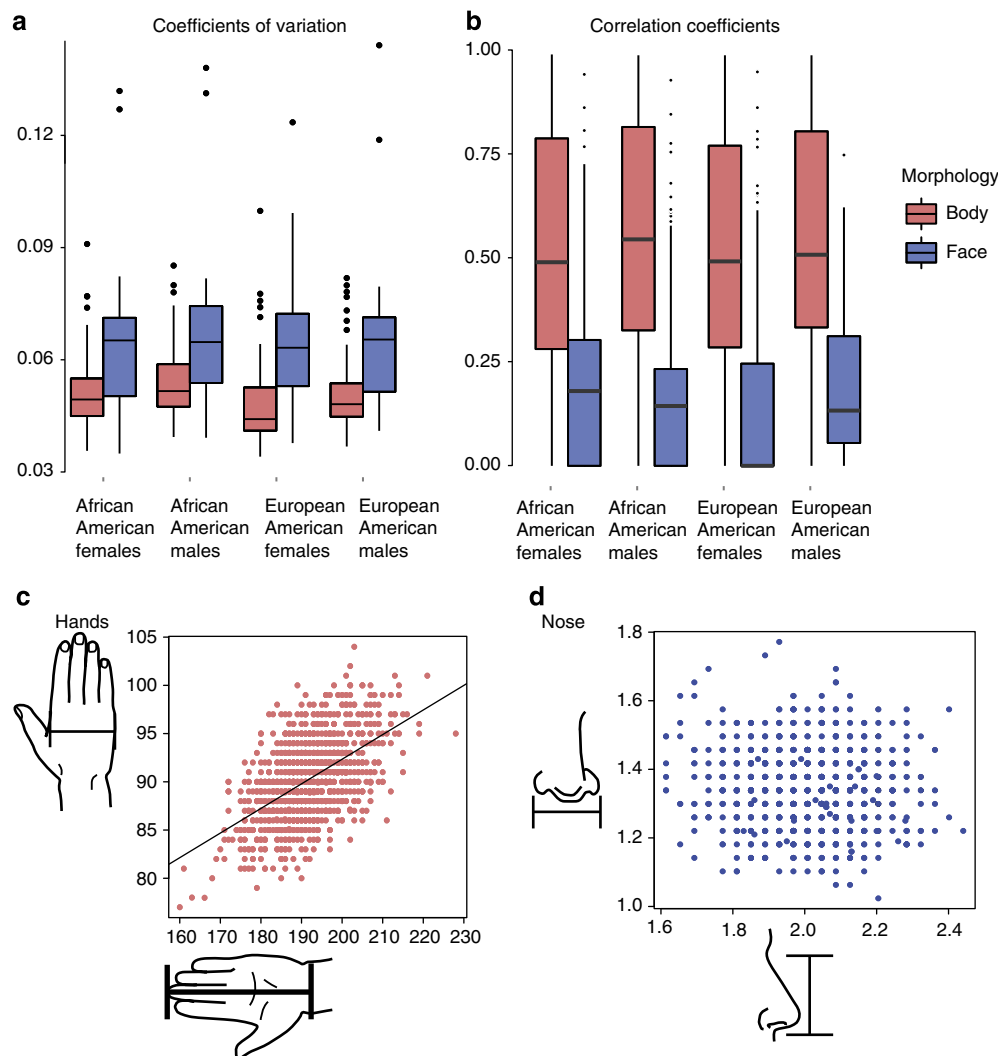


Figure 2 | Morphological evidence that human faces have evolved to signal individual identity. Morphological comparisons of facial features with other aspects of body morphology are consistent with selection for identity signals. **(a)** In all four groups examined, facial traits have higher coefficients of variation than other body traits ($P < 0.03$ for all comparison). **(b)** Facial traits as a group show lower inter-trait correlations than non-facial traits in all four populations examined ($P < 0.001$ for all comparisons). **(c)** For most traits, such as hands, larger individuals have larger traits such that the width and length of an individual's hand are correlated. **(d)** In contrast to hands, the width and length of the nose are not correlated. Box plots show median and 25th and 75th percentiles ($N = 181$ African American females; 457 African American males; 204 European American females; 1,168 European American males). The P -values shown in the figure legend are from one-tailed Mamm-Whitney U tests. The scatterplots show the trait values for European American male service members measured in the ANSUR II data set. Best-fit lines are shown for significant regressions.

different populations could maintain diversity at distinct loci underlying the same trait; or selection may act on loci underlying different traits in each population depending on the dynamics of selection as human populations expanded across the globe. We explored this question by assessing whether loci showing elevated diversity, where both π corrected for divergence with macaques and Tajima's D both fall in the 95th percentile, were shared across populations. Indeed, a disproportionate number of loci show evidence of elevated diversity in at least one population for faces (9/58) compared with height (6/356; $\chi^2 = 23.5$, $P < 0.0001$) and intergenic regions (57/4,873; $\chi^2 = 78.8$, $P < 0.0001$, Fig. 4). Furthermore, the regions that show elevated diversity for faces are more consistent across continents than expected; five of nine regions show elevated diversity on at least two continents compared with 1 of 57 intergenic regions ($\chi^2 = 27.2$, $P < 0.0002$, Fig. 4). All nine regions identified as having elevated diversity in at least one population have high levels of π /divergence (> 90 th percentile) in both African

populations examined here (Supplementary Table S3). Additionally, analyses of the haplotype networks for the nine regions show greater allelic diversity in African populations, with European and Asian populations carrying a subset of the African haplotypes (Supplementary Figs 11–19). These patterns of diversity and haplotype sharing across populations are consistent with an African origin of allelic variation at identity signalling loci predating human migration out of Africa. Population differentiation in facial morphology appears in part to be the result of differential loss of diversity in non-African populations, consistent with reduced morphological variation in populations with increased distance to Africa⁴⁵.

Here we present two examples highlighting the complex evolutionary trajectories of loci involved in identity signalling in humans. The examples illustrate (i) that genetic variation underlying identity signals tends to be old and of African origin and (ii) that phenotypic divergence between non-African populations is partly related to the differential loss of ancestral

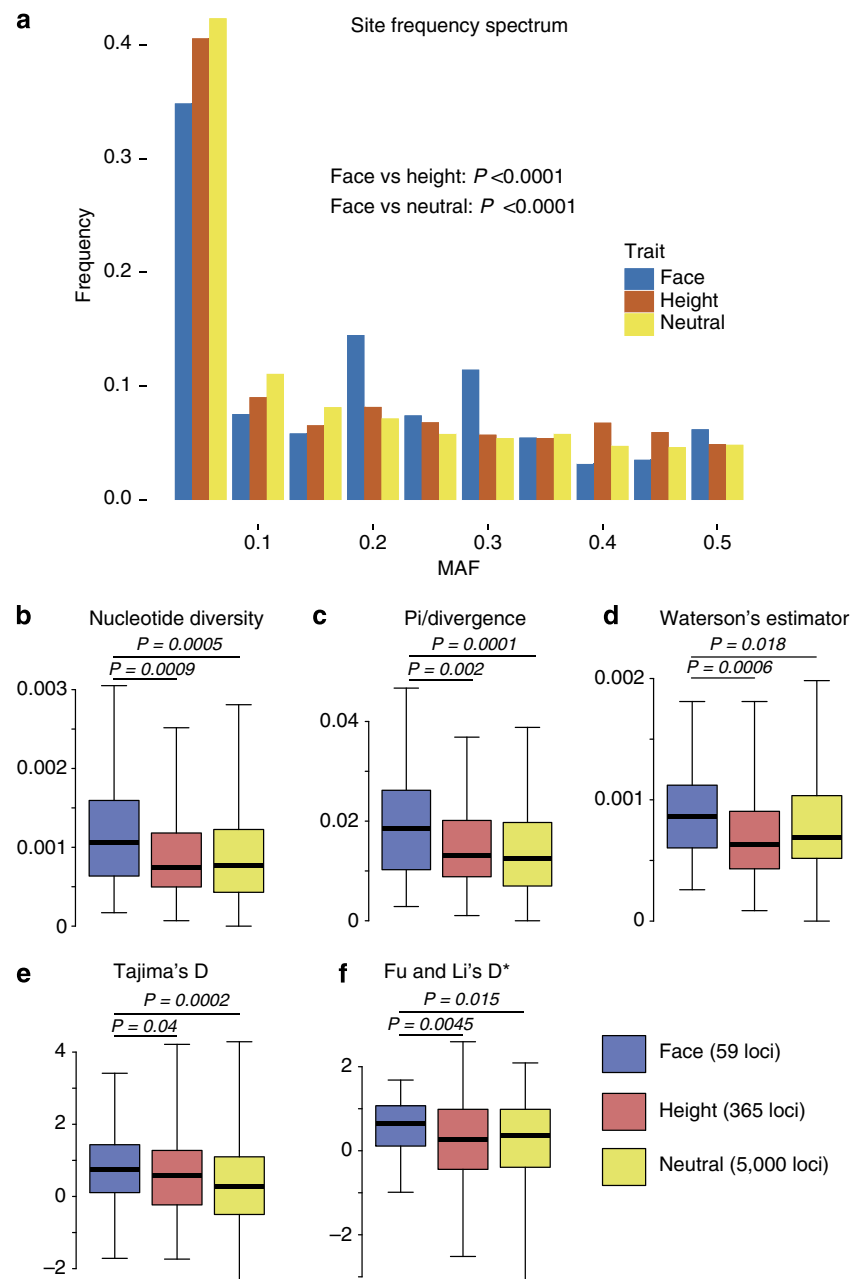


Figure 3 | Population genomic evidence that human faces have evolved to signal individual identity. Genomic regions associated with facial morphology show evidence of selection for identity signalling in the Finnish. **(a)** Face regions ($N = 59$) have elevated levels of intermediate-frequency alleles compared with neutral regions ($N = 5,000$) or genomic regions associated with variation in height ($N = 365$). The bar graph shows the proportion of SNPs within each minor allele frequency (MAF) bin. **(b)** Additionally, face regions have elevated levels of π , **(c)** even after controlling for differing rates of divergence among loci. **(d)** Similarly, face regions show an elevated number of segregating sites, measured as Watterson's θ . **(e)** Tajima's D and **(f)** Fu and Li's D* are elevated in facial regions compared to height and neutral intergenic controls regions. Whiskers shows the 5th and 95th percentiles. Outliers are not shown so that the main distributions can be viewed at larger size. The P -values shown are from one-tailed Mann-Whitney U -tests. Note that sample sizes are reduced for tests corrected for divergence, as alignments were not available for all regions considered ($N = 58$ face loci, 356 height loci, 4,873 neutral loci).

variation (Fig. 5). Variants associated with the distance between the chin and bridge of the nose³³ are found within an intron of *TMTC2*. A sliding window analysis of the region demonstrates that there is elevated diversity and reduced *Fst* consistent with sustained selection for identity signalling that is common to the three continental groups or occurred in their ancestral population (Fig. 5a). In contrast to the shared diversity at *TMTC2*, intronic variants of *SDK1* associated with nasal morphology³² show a clear reduction of nucleotide diversity in Asian populations compared

with the elevated diversity found in African populations. The reduction in diversity in Asian populations and increased *Fst* between Asian and African populations could either be the result of loss of diversity during population bottlenecks or directional selection on nasal morphology in Asian populations (Fig. 5b). For both loci, we constructed gene trees for the 5-kb window with the highest level of nucleotide diversity for 30 modern human sequences as well as the Neanderthal, Denisovan and chimpanzee sequences (Fig. 5c,d). Both trees provide further evidence for the

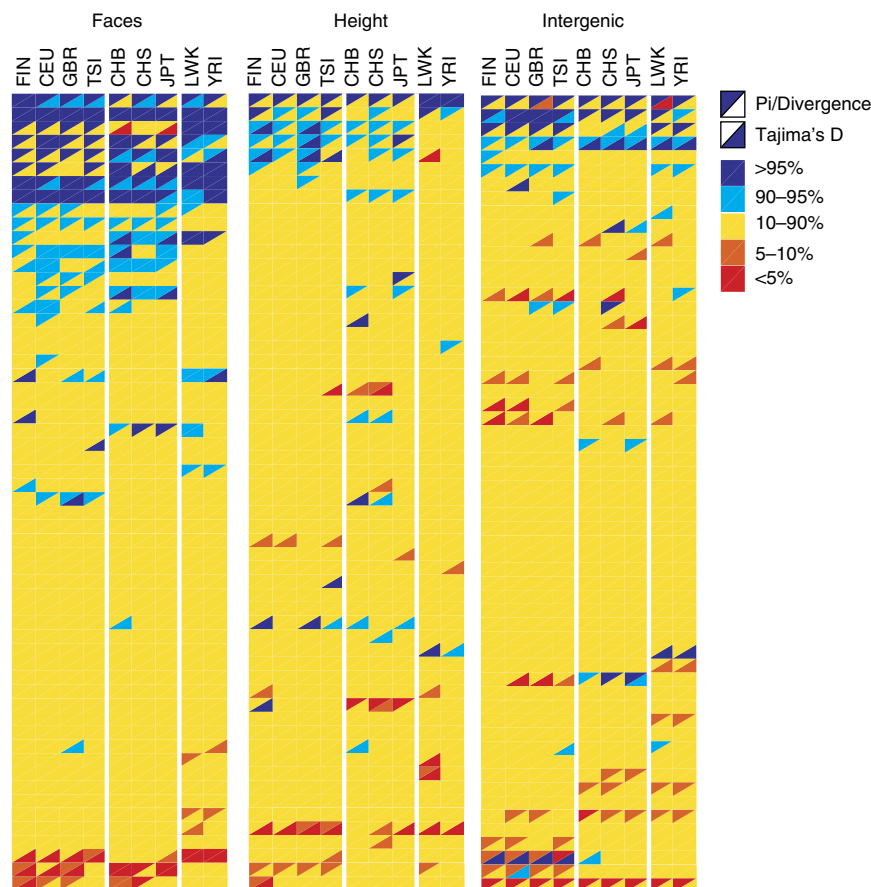


Figure 4 | Patterns of elevated diversity in face-associated loci across populations. The face-associated loci with elevated diversity consistent with selection for identity signalling tend to be shared across populations both within and between continents. The heatmap highlights loci on the extreme ends of the distributions for π (controlling for divergence with macaque) and Tajima's D. Columns correspond to populations and rows correspond to individual loci. Squares that are fully filled in with dark blue designate loci with evidence of elevated diversity (>95th percentile for both summary statistics). A greater number of loci show evidence of elevated diversity in at least one population for faces (9/58) compared with height (6/356; $\chi^2 = 23.5$, $P < 0.0001$) and intergenic regions (57/4,873; $\chi^2 = 78.8$, $P < 0.0001$). Additionally, patterns of elevated diversity are more consistently shared across populations for face-associated regions compared with the neutral regions (5/9 face regions versus 1/57 neutral regions, $\chi^2 = 27.2$, $P < 0.0002$). To facilitate visual comparison, representative subsamples of height and intergenic regions are shown here. Subsamples were generated by randomly selecting loci from the height and neutral lists, which we confirmed did not deviate from the distribution of the total sample. All analyses reported were conducted on the full data sets.

ancient origins of loci under selection for identity signalling as archaic Hominin species are nested within modern human diversity. This result suggests that selection on some loci associated with identity signalling predates the origin of *Homo sapiens* and the emergence of modern facial morphology.

Discussion

Here we have presented both morphological and population genomic evidence consistent with the hypothesis that selection for individual identity signals has shaped patterns of human facial diversity. Though the evidence for selection at individual loci is modest, as expected for molecular evolution of polygenic traits³⁵, the combination of morphological and genomic data from multiple populations clearly rejects the identity cue hypothesis and provides compelling evidence consistent with the idea that selection for individual identity signalling has shaped patterns of facial morphology in humans. Provided that the variation used in identity signals is not developmentally costly to produce or maintain, even a small selective advantage of individuality is expected to give rise to elevated phenotypic diversity when confusion is costly²¹. Previous studies have shown that being

confused with others may be costly in a range of circumstances including within social hierarchies in *Polistes* wasps²⁶, sexual selection in house mice²⁷ and parent-offspring interactions in cliff swallows²⁸. It is unknown at present, which aspects of human sociality have been the most important sources of selection for identity signalling though it is likely that multiple facets of social interactions contribute to selection for identity signals. Individual recognition and discriminating among individuals plays a role in shaping important human behaviours including kin recognition⁴⁶, investment in offspring⁴⁷ and cooperation⁴⁸. It is likely that many social contexts favour identity signalling in humans, so it will be important for future research to explore the relative benefits of individuality across many social contexts and developmental stages in humans.

In addition to selection for identity signalling, it is possible that other frequency-dependent process such as preferences for mates with rare or novel features could have played a role in shaping human diversity. For example, a recent study showed frequency-dependent effects on the attractiveness of male facial hair styles¹⁹. Preferences for individuals with rare phenotypes have also been shown in other animals, such as guppies where rare phenotypes confer a survival advantage due to reduced predation^{14,17}. In

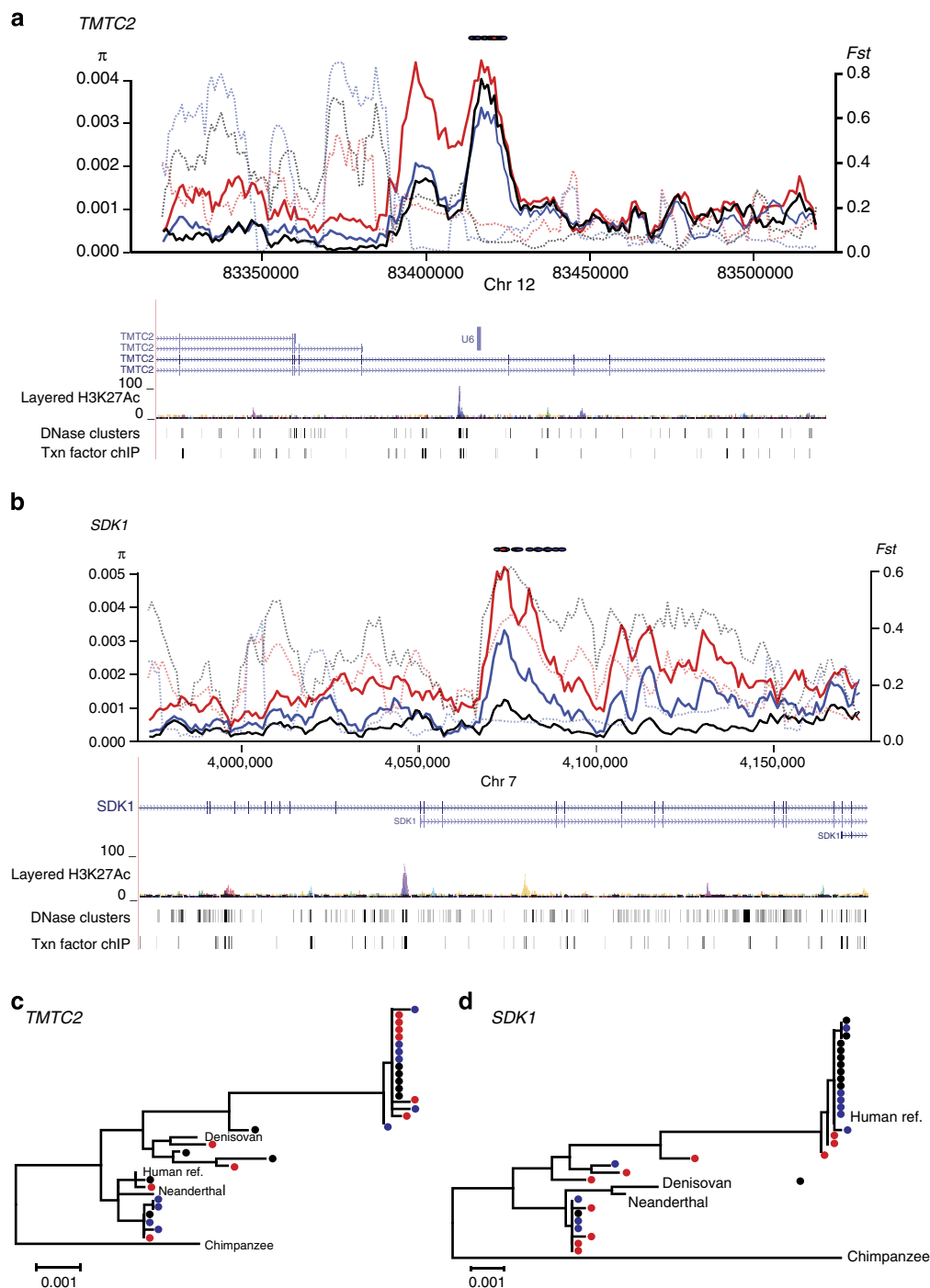


Figure 5 | Evolutionary history of example face-associated loci. Patterns of genetic diversity associated with facial morphology at *TMT2* and *SDK1*. **(a)** At *TMT2* variation is largely shared across continents, while **(b)** at *SDK1* variation has been lost mainly in the CHB population. The sliding window analyses **(a,b)** show nucleotide diversity for three 1000 Genomes populations representing Europe (FIN), Asia (CHB) and Africa (YRI), respectively for 5-kb windows at 1-kb sliding intervals. Nucleotide diversity is shown with solid lines while F_{st} is represented by dotted lines. Colour of the lines represents the population examined for π (FIN = blue, CHB = black, YRI = red) or the two population F_{st} comparisons (FIN-YRI = red, CHB-YRI = black, FIN-CHB = blue). The locations of SNPs associated with facial morphology are shown as blue circles except for the focal SNP included in other window-based analyses that is denoted with a red circle. The UCSC Genome Browser tracks showing the locations of exons and three ENCODE regulatory regions, which show regions likely associated with genomic features involved in gene regulation, are shown below the sliding window. **(c,d)** Maximum likelihood trees show the relationships among 10 modern humans sampled from each of three populations (FIN, CHB and YRI) as well as sequences from Denisovan, Neanderthal and Chimpanzee. The modern human sequences are coloured according to their population of origin (FIN = blue, CHB = black, YRI = red). The region analyzed was the 5-kb window with the highest nucleotide diversity as determined by the sliding window analysis. Note that in both cases, the sequences for archaic Hominins are nested within modern human diversity, indicating the origin of the major haplogroups predates the evolution of *Homo sapiens*.

humans, females tend to advertise physical attractiveness to mate more prominently than do males, who tend to advertise resources or performance ability^{49,50}. Thus if frequency-dependent mate preferences were the major driver in determining facial identity, then females might be expected to show elevated levels of individuality compared with males just as female preferences for novel individuals contributes to the elevated colour pattern variation seen in male guppies¹⁷. However, in humans both males and females show elevated individuality in faces compared with other external morphology (Fig. 2), suggesting that mate preferences alone cannot explain the patterns observed here. Similarly, mate preferences might be expected to drive variation only in adults as is observed in guppies¹⁷, yet distinctive facial morphology is seen at all life stages in humans. To the extent that frequency-dependent mate preferences play a role in shaping patterns of facial individuality, it is likely that mating preferences and identity signalling would have a positive feedback. If individual distinctiveness is beneficial in non-sexual contexts, preferences for mates with rare phenotypes may then also provide an additional benefit to distinctiveness¹⁵. Finally, our data do not preclude potential directional or stabilizing selection pressures that may arise from other potential mating preferences⁵¹ or climate⁵² on particular features of human facial morphology, though directional and stabilizing selection do not predict elevated genetic diversity within populations at the associated loci and so cannot explain the patterns of elevated genetic variation we have documented here.

It is important to note that facial recognition is widespread in primates⁵³ and identity signalling is unlikely to be limited to human facial morphology, though the loci under selection may vary considerably across species. This may be especially true for humans, which have undergone considerable directional evolution of facial form during the course of hominin evolution⁵⁴. While faces are a key feature used in human social recognition, other traits such as our voices also contribute to recognition and may have also experienced selection for identity signalling. Additionally, the strength of selection for particular identity-signalling traits may have changed over time in modern humans as cultural practices gave rise to individually distinctive clothing and hairstyles, which provide additional cues to identity. Traditional treatments of social selection in human evolution have emphasized the potential role for social interactions in shaping our cognitive abilities³, though our work demonstrates that social selection has shaped our morphology as well to facilitate social recognition. Importantly, our work draws a link between social interactions and the maintenance of genetic variation underlying traits used in social recognition. Social recognition is found across disparate animal taxa suggesting that selection for identity signalling is likely to be a common mechanism generating phenotypic variation and maintaining genetic variation.

Methods

Morphological analyses. We examined morphological relationships among body parts and facial features using published anthropometric data sets. We focused our analyses on the ANSUR II data set because it provides a large, consistent database of individual level facial and body measurements. We analyzed the linear anthropometric measurements (Supplementary Tables 1 and 2). In our analysis, we considered four groups of service members based on their sex and racial identity: African American females ($n = 181$, mean height = 64.29 ± 0.17 inches), African American males ($n = 457$, mean height = 69.12 ± 0.13 inches), European American females ($n = 204$, mean height = 64.27 ± 0.15 inches), and European American males ($n = 1,168$, mean height = 69.32 ± 0.08 inches). Compared with the general civilian population, the individuals measured in the ANSUR II data set tend to be taller and have lower levels of body fat⁵⁵. Neither of these factors should influence our results or conclusions because our comparisons use facial and body measurements from the same individuals. Identity signalling predicts that traits used for recognition will have greater variance and be less correlated with each other compared with non-recognition traits in the same group of individuals.

Using the ANSUR II data set, we tested two predictions of the identity-signalling hypothesis. First, we considered the levels of variation in each trait by calculating the coefficient of variation—by dividing the standard deviation of each trait by the mean. Coefficients of variation provide a scale-free method for comparing variation across samples that differ in average size as is the case for human morphological data. Second, we considered the correlations among traits by calculating the inter-trait Pearson's correlations for all pairwise combination of traits within each class of traits. To compare the distribution of correlation coefficients between bodies and faces, we recorded the correlation coefficients significant at the $P < 0.05$ level. For any pair of traits which did not show a significant correlation at $P < 0.05$, we recorded the correlation as 0. Pearson's correlation test is sensitive to the sample size such that correlations are more likely to be significant when larger samples are used. Therefore, comparisons between the different groups considered should be made with caution because of differences in sample size. For example, differences in the percentage of significant pairwise comparisons between males and females likely reflect differences in samples sizes. Within a group, however, the same individuals were measured for both facial and body traits, providing a direct comparison of the relative degree of correlation among traits.

Selection of genomic regions for analysis. Face-associated SNPs were taken from two recent genome-wide association studies of normal facial morphology. Paternoster *et al.*⁵³ conducted a discovery phase association study where they examined the relationship between facial characteristics and more than 2.5 million imputed SNPs in a sample of 2,185 15 year olds from the Avon Longitudinal Study of Parents and their Children⁵⁶. Only subjects who genetically clustered with the CEU HapMap population were included in their analysis. The study identified 30 loci associated with facial morphology at $P < 5 \times 10^{-7}$, which we examined in our study. Liu *et al.*³² examined the relationship between facial morphology and more than 2.5 million SNPs in a discovery phase association study of 5,388 adults. The samples in the Liu *et al.*³² study came individuals of European ancestry living in the Netherlands, Australia, Canada, Germany and the United Kingdom. They identified 29 loci associated with facial morphology at $P < 5 \times 10^{-7}$, which we examined in this study. None of the SNPs identified by the two studies overlapped, providing a total of 59 loci for investigation. In both studies, multiple linked SNPs were often identified in association with a particular phenotype. When more than one SNP was associated with a trait, we chose the SNP with the smallest P -value within a 1-MB region of a chromosome from the association study. The SNPs identified for further examination from the two studies include one from each of 59 loci distributed throughout the autosomes. The SNPs are largely intergenic (95%) though a few occur within introns (5%). None were located in coding regions.

We compared face-associated genomic regions with two sets of control regions. First, we examined SNPs associated with height taken from the genome-wide association study catalog of the National Human Genome Research Institute (www.genome.gov) on 25 April 2013. In order to prevent multiple sampling of any regions, we only considered SNPs that were separated by more than 4 kb. In the instances where multiple nearby SNPs had been associated with height, we chose the SNP that had been associated with the smallest P -value as reported in the genome-wide association study catalog. We excluded six SNPs associated with height that fall within the HLA region, though including the SNPs in our analyses does not alter our pattern of results. This produced a total of 365 loci associated with variation in height. Like faces, height is a composite character that depends on the morphology of numerous different bones. Additionally, both height and facial morphology have complex genetic bases with many loci of small effect contributing overall phenotypic variation⁴³. SNPs associated with height are predominantly located in intronic regions (54%) and intergenic regions (36%) with a smaller percent found near the 3' and 5' end of genes (6%) or exons (3%). Second, we considered the genome-wide patterns of diversity by examining 5,000 2 kb intergenic regions. We identified putatively neutral intergenic regions in Europeans using the Neutral Region Explorer webserver⁴². The same set of intergenic regions was used for all populations.

We analyzed regions surrounding the SNPs identified by the association studies at a set window size. The causative mutations underlying the traits are not known, though are likely to be located near the SNPs identified through genome-wide association studies⁵⁷. The *a priori* best choice for a window size is not clear, though the patterns of elevated nucleotide diversity we observe are seen over a range of window sizes (Supplementary Fig. 1). We chose to analyze 1 kb both up and downstream of the SNPs, providing windows of 2 kb. Smaller window sizes show marked increased variance in the summary statistics across loci (Supplementary Fig. 1), though this variance levels off at window sizes of 2 kb or greater.

Summary statistics. We calculated summary statistics for each population using binary SNP and indel data from 1000 Genome Project Phase 1 variants. Nine non-admixed populations originating from Europe (CEU: Utah residents with Northern and Western European Ancestry; GBR: British from England and Scotland; FIN: Finnish from Finland; TSI: Toscani from Italy), Asia (CHB: Han Chinese in Beijing, China, CHS: Southern Han Chinese, JPT: Japanese from Tokyo, Japan) and Africa (LWK: Luhya from Webuye, Kenya; YRI: Yoruba from Ibadan, Nigeria) were considered in our study. We downloaded the population data to Galaxy⁵⁸ using the Table Browser function of the UCSC genome browser. We filtered the

data based on the sets of 2-kb windows for face, height and intergenic regions to produce three files for each population, which we subsequently examined using custom macros in Excel.

We used folded site frequency spectra to examine the distribution of minor allele frequencies among SNPs found within each of the demarcated regions. The expected distribution of allele frequencies at loci underlying a polygenic trait under negative frequency-dependent selection is unclear and will depend on the exact form of selection and the genetic architecture of the trait⁵⁹. Nonetheless, frequency-dependent selection is expected to maintain alleles in a population, on average, longer than expected for neutral alleles⁶⁰. Thus, the distribution of allele frequencies should differ from that expected in a stationary population at mutation drift equilibrium. In particular, we expect fewer rare alleles under a scenario of frequency-dependent selection. Spectra were compared using the raw counts of SNPs with each minor allele frequency using a MWU-test. To graph the folded site frequency spectra, we binned data into ranges of minor allele frequencies.

In addition to the aggregated site frequency spectrum analysis, we also considered the distribution of multiple summary statistics of genetic diversity across the loci considered within our study. We calculated the following summary statistics for each 2-kb window: π , π corrected for human-macaque divergence, Watterson's θ , Tajima's D and Fu and Li's D^* . Both π and θ are estimators of the neutral mutation parameter, $4N_e\mu$. π is based on the number of pairwise differences among sequences within a sample and θ is based on the proportion of segregating sites. Loci under frequency-dependent selection are expected to show elevated values for π because frequency-dependent selection maintains alleles over longer periods of time. Older alleles accumulate mutations and therefore show higher levels of pairwise sequence divergence. We also examined the distribution of π corrected for the rate of divergence between humans and macaques. Different regions of the genome are known to experience differences in rates of mutation⁶¹. Loci with higher mutation rates will show elevated levels of π . The rate of divergence between humans and macaques provides a means of estimating the relative differences in mutation rates among loci⁶². The maintenance of multiple alleles in a population under frequency-dependent selection is also expected to lead to higher estimates of θ . Tajima's D is the normalized difference between π and θ . Tajima's D takes on positive values when there is an excess of intermediate-frequency variants and negative values when there is an excess of rare variants. Fu and Li's D^* is based on the number of nucleotide variants observed only once in a sample⁶³. Negative measures of Fu and Li's D^* indicate an excess of singletons. Loci under frequency-dependent selection are expected to have a relatively smaller number of singletons and therefore more positive values of Fu and Li's D^* .

We calculated the summary statistics using the allele frequencies given in the phase 1 variant files from the 1000 Genomes project. The short indels recorded in the data set were considered in the same manner as SNPs. Human-macaque divergence data were estimated using the LastZ alignment of the two reference genomes. Only regions with alignments between the two species' genomes were considered in the analysis of π corrected for divergence with macaques (faces = 58 regions, height = 356 regions, neutral = 4,873 regions) and for subsequent analyses using this statistic. For the aligned regions, the average alignment lengths were $1,832.21 \pm 4.49$ sites out of 2,000. We compared the distribution of summary statistics for face-associated loci with the distributions for the two control data sets using one-tailed MWU-tests.

Patterns of diversity across populations. We asked whether the same loci showed elevated diversity in different populations. To do this, we identified loci in each population for which π /divergence and Tajima's D were above the 95th percentile as determined from the empirical distribution of intergenic regions examined within that population. We then asked whether or not a disproportionate number of loci with elevated diversity was shared between continents for face regions compared with the intergenic regions examined. For the nine loci showing elevated diversity in at least one population, we investigated the patterns of haplotype sharing across populations. We examined the sequences in the 2-kb window used for previous analyses. For those coordinates, we downloaded a combined PED file including CHB, FIN and YRI from the 1000 Genomes project site (browser.1000genomes.org). We converted the PED files to fasta format using PGD Spider⁶⁴. This procedure produced a fasta file containing the polymorphic sites found within the examined loci. Using the 'pegas' package in R⁶⁵, we created haplotype networks for each of the loci.

Sliding window analyses. To examine the extent to which selection for identity signalling has been shared or divergent across continents, we conducted a sliding window analysis of the regions identified as having elevated diversity in at least one population. We calculated π and F_{st} for 5-kb windows every 1 kb for a total of 200 kb. π was calculated for one representative population for each continent (FIN, CHB and YRI). We estimated levels of differentiation between populations using Hudson's F_{st} following ref. 66 as it produces unbiased estimates of F_{st} and is less sensitive to sample size and rare variants than other estimates of F_{st} such as Weir and Hill⁶⁷. We estimated F_{st} for each set of SNPs considered by calculating a ratio of averages rather than an average of ratios, as the former is less sensitive to the presence of rare variants in a sample⁶⁶.

The SNPs identified in association with facial morphology are not found in coding sequences, so they are likely to influence gene regulation or splicing in some manner. For the two loci examined in greater detail, we used the UCSC genome browser to identify polymorphic sites in ENCODE regulatory regions. We focused on three ENCODE tracks in the UCSC browser⁶⁸: H3K27Ac marks, DNase sensitivity clusters and transcription factor-binding sites. The H3K27Ac marks show regions for which there is CHIP-seq based evidence of enrichment for the H3K27Ac histone mark. H3K27 acetylation is associated with enhanced transcription. DNase sensitivity clusters show regions sensitive to DNase as assessed across 125 cell types. Promoters and other regulatory regions tend to be DNase sensitive. The transcription factor track shows regions with evidence of transcription factor-binding sites.

Gene trees. For the two 5-kb loci examined, we constructed maximum likelihood gene trees with 10 sequences each from the FIN, CHB and YRI 1000 Genomes populations for a total of thirty sequences. Additionally, we included the human and chimpanzee reference sequences as well as sequences for Denisovans⁶⁹ and Neanderthals (<http://cdna.eva.mpg.de/neandertal/altai/> AltaiNeandertal/bam/). We downloaded the alignment of the human and chimpanzee reference sequences from Ensembl. Denisovan sequences were downloaded using the Table Browser function of the UCSC Genome Browser. The draft Altai Neanderthal sequences were downloaded for the relevant chromosomes from the Department of Evolutionary Genetics at the Max Planck Institute's website. We constructed individual sequences for the 1000 Genomes, Denisovan and Neanderthal by manually altering the human reference sequence in accordance with the data found in the respective VCF files using Mega 5.2.1 (ref. 70). For the phased 1000 Genomes data, we selected one chromosome per individual sample. For the Neanderthal and Denisovan sequences, we included all of the sites that differed from the human reference to make a single sequence. After removing sites with gaps in the alignment, we constructed a maximum likelihood tree using a general time reversible model with a gamma distribution of invariant sites.

References

- Dunbar, R. I. M. Social cognition on the internet: testing constraints on social network size. *Philos. Trans. R. Soc. B Biol. Sci.* **367**, 2192–2201 (2012).
- Apicella, C. L., Marlowe, F. W., Fowler, J. H. & Christakis, N. A. Social networks and cooperation in hunter-gatherers. *Nature* **481**, 497–501 (2012).
- Dunbar, R. I. M. & Shultz, S. Evolution in the social brain. *Science* **317**, 1344–1347 (2007).
- Byrne, R. W. & Whiten, A. *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans* (Oxford University Press, 1988).
- Whiten, A. & Erdal, D. The human socio-cognitive niche and its evolutionary origins. *Philos. Trans. R. Soc. B Biol. Sci.* **367**, 2119–2129 (2012).
- Haxby, J. V., Hoffman, E. A. & Gobbini, M. I. Human neural systems for face recognition and social communication. *Biol. Psychiatry* **51**, 59–67 (2002).
- Herrmann, E., Call, J., Hernandez-Lloreda, M. V., Hare, B. & Tomasello, M. Humans have evolved specialized skills of social cognition: the cultural intelligence hypothesis. *Science* **317**, 1360–1366 (2007).
- Parr, L. A. The evolution of face processing in primates. *Philos. Trans. R. Soc. B Biol. Sci.* **366**, 1764–1777 (2011).
- Pascalis, O. & Kelly, D. J. The origins of face processing in humans: phylogeny and ontogeny. *Perspect. Psychol. Sci.* **4**, 200–209 (2009).
- Kanwisher, N. & Yovel, G. The fusiform face area: a cortical region specialized for the perception of faces. *Philos. Trans. R. Soc. B Biol. Sci.* **361**, 2109–2128 (2006).
- Light, L. L., Kayra-Stuart, F. & Hollander, S. Recognition memory for typical and unusual faces. *J. Exp. Psychol. Hum. Learn.* **5**, 212 (1979).
- Tibbetts, E. A. & Dale, J. Individual recognition: it is good to be different. *Trends Ecol. Evol.* **22**, 529–537 (2007).
- Bond, A. B. & Kamil, A. C. Apostatic selection by blue jays produces balanced polymorphism in virtual prey. *Nature* **395**, 594–596 (1998).
- Olendorf, R. *et al.* Frequency-dependent survival in natural guppy populations. *Nature* **441**, 633–636 (2006).
- Kokko, H., Jennions, M. D. & Houde, A. Evolution of frequency-dependent mate choice: keeping up with fashion trends. *Proc. R. Soc. B Biol. Sci.* **274**, 1317–1324 (2007).
- Dale, J. in *Bird Coloration Volume 2 Function and Evolution* (Harvard University Press, 2006).
- Hughes, K. A., Houde, A. E., Price, A. C. & Rodd, F. H. Mating advantage for rare males in wild guppy populations. *Nature* **503**, 108–110 (2013).
- Frost, P. European hair and eye color: a case of frequency-dependent sexual selection? *Evol. Hum. Behav.* **27**, 85–103 (2006).
- Janif, Z. J., Brooks, R. C. & Dixon, B. J. Negative frequency-dependent preferences and variation in male facial hair. *Biol. Lett.* **10**, 20130958 (2014).
- Johnstone, R. A. Recognition and the evolution of distinctive signatures: when does it pay to reveal identity? *Proc. R. Soc. Lond. Ser. B Biol. Sci.* **264**, 1547–1553 (1997).

21. Dale, J., Lank, D. B. & Reeve, H. K. Signaling individual identity versus quality: a model and case studies with ruffs, queleas, and house finches. *Am. Nat.* **158**, 75–86 (2011).
22. Scott-Phillips, T. C. Defining biological communication. *J. Evol. Biol.* **21**, 387–395 (2008).
23. Thom, M. D. & Hurst, J. L. Individual recognition by scent. *Ann. Zool. Fenn.* **41**, 765–787 (2004).
24. Bergman, T. J. & Sheehan, M. J. Social knowledge and signals in primates. *Am. J. Primatol.* **75**, 683–694 (2013).
25. Sheehan, M. J. & Tibbetts, E. A. Selection for individual recognition and the evolution of polymorphic identity signals in *Polistes* paper wasps. *J. Evol. Biol.* **23**, 570–577 (2010).
26. Sheehan, M. J. & Tibbetts, E. A. Evolution of identity signals: frequency-dependent benefits of distinctive phenotypes used for individual recognition. *Evolution* **63**, 3106–3113 (2009).
27. Thom, M. D. F. & Dytham, C. Female choosiness leads to the evolution of individually distinctive males. *Evolution* **66**, 3736–3742 (2012).
28. Medvin, M. B., Stoddard, P. K. & Beecher, M. D. Signals for parent offspring recognition: a comparative analysis of the begging calls of cliff swallows and barn swallows. *Anim. Behav.* **45**, 841–850 (1993).
29. Pollard, K. A. & Blumstein, D. T. Social group size predicts the evolution of individuality. *Curr. Biol.* **21**, 413–417 (2011).
30. Bamshad, M. & Wooding, S. P. Signatures of natural selection in the human genome. *Nat. Rev. Genet.* **4**, 99–111 (2003).
31. Andrés, A. M. *et al.* Targets of balancing selection in the human genome. *Mol. Biol. Evol.* **26**, 2755–2764 (2009).
32. Liu, F. *et al.* A genome-wide association study identifies five loci influencing facial morphology in Europeans. *PLoS Genet.* **8**, e1002932 (2012).
33. Paternoster, L. *et al.* Genome-wide association study of three-dimensional facial morphology identifies a variant in *PAX3* associated with nasion position. *Am. J. Hum. Genet.* **90**, 478–485 (2012).
34. Attanasio, C. *et al.* Fine tuning of craniofacial morphology by distant-acting enhancers. *Science* **342**, 1241006 (2013).
35. Pritchard, J. K., Pickrell, J. K. & Coop, G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr. Biol.* **20**, R208–R215 (2010).
36. Biswas, S. & Akey, J. M. Genomic insights into positive selection. *Trends Genet.* **22**, 437–446 (2006).
37. Gordon, C. C., Churchill, T., Clauser, C. E., Bradtmiller, B. & McConville, J. T. *Anthropometric Survey of US Army Personnel: Methods and Summary Statistics 1988* (DTIC Document, 1989).
38. Frankino, W. A., Zwaan, B. J., Stern, D. L. & Brakefield, P. M. Natural selection and developmental constraints in the evolution of allometries. *Science* **307**, 718–720 (2005).
39. Huxley, J. *Problems of Relative Growth* (1932).
40. Marroig, G., Shirai, L. T., Porto, A., de Oliveira, F. B. & De Conto, V. The evolution of modularity in the mammalian skull II: evolutionary consequences. *Evol. Biol.* **36**, 136–148 (2009).
41. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **473**, 544–544 (2011).
42. Arbiza, L., Zhong, E. & Keinan, A. NRE: a tool for exploring neutral loci in the human genome. *BMC Bioinformatics* **13**, 301 (2012).
43. Allen, H. L. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).
44. Barrett, J. C. & Cardon, L. R. Evaluating coverage of genome-wide association studies. *Nat. Genet.* **38**, 659–662 (2006).
45. Manica, A., Amos, W., Balloux, F. & Hanihara, T. The effect of ancient population bottlenecks on human phenotypic variation. *Nature* **448**, 346–348 (2007).
46. Lieberman, D., Tooby, J. & Cosmides, L. The architecture of human kin detection. *Nature* **445**, 727–731 (2007).
47. Fernandez-Duque, E., Vagstad, C. R. & Mendoza, S. P. The biology of paternal care in human and nonhuman primates. *Annu. Rev. Anthropol.* **38**, 115–130 (2009).
48. Brosnan, S. F., Salwiczek, L. & Bshary, R. The interplay of cognition and cooperation. *Philos. Trans. R. Soc. B Biol. Sci.* **365**, 2699–2710 (2010).
49. Anderson, R. C. & Klostad, C. A. For love or money? the influence of personal resources and environmental resource pressures on human mate preferences. *Ethology* **118**, 841–849 (2012).
50. Bereczkei, T., Voros, S., Gal, A. & Bernath, L. Resources, attractiveness, family commitment; reproductive decisions in human mate choice. *Ethology* **103**, 681–699 (1997).
51. Puts, D. A., Jones, B. C. & DeBruine, L. M. Sexual selection on human faces and voices. *J. Sex Res.* **49**, 227–243 (2012).
52. Hubbe, M., Hanihara, T. & Harvati, K. Climate signatures in the morphological differentiation of worldwide modern human populations. *Anat. Rec.* **292**, 1720–1733 (2009).
53. Parr, L. A. The evolution of face processing in primates. *Philos. Trans. R. Soc. B Biol. Sci.* **366**, 1764–1777 (2011).
54. Wood, B. & Harrison, T. The evolutionary context of the first hominins. *Nature* **470**, 347–352 (2011).
55. Fromuth, R. & Parkinson, M. Predicting 5th and 95th percentile anthropometric segment lengths from population stature. *Proc. ASME Int. Des. Eng. Tech. Conf.* 3–6 (2008).
56. Golding, J., Pembrey, M., Jones, R. & Team, A. S. ALSPAC-the avon longitudinal study of parents and children. I. study methodology. *Paediatr. Perinat. Epidemiol.* **15**, 74–87 (2001).
57. Orozco, G., Barrett, J. C. & Zeggini, E. Synthetic associations in the context of genome-wide association scan signals. *Hum. Mol. Genet.* **19**, R137–R144 (2010).
58. Blankenberg, D. *et al.* Galaxy: a web-based genome analysis tool for experimentalists. *Curr. Protoc. Mol. Biol.* **10**, 1–21 (2010).
59. Navarro, A. & Barton, N. H. The effects of multilocus balancing selection on neutral variability. *Genetics* **161**, 849–863 (2002).
60. Takahata, N. & Nei, M. Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. *Genetics* **124**, 967–978 (1990).
61. Wolfe, K. H., Sharp, P. M. & Li, W.-H. Mutation rates differ among regions of the mammalian genome. *Nature* **19**, 283–285 (1989).
62. Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111–120 (1980).
63. Fu, Y. X. & Li, W. H. Statistical tests of neutrality of mutations. *Genetics* **133**, 693–709 (1993).
64. Lischer, H. E. L. & Excoffier, L. PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics* **28**, 298–299 (2012).
65. Paradis, E. pegas: an R package for population genetics with an integrated-modular approach. *Bioinformatics* **26**, 419–420 (2010).
66. Bhatia, G., Patterson, N., Sankaraman, S. & Price, A. L. Estimating and interpreting FST: the impact of rare variants. *Genome Res.* **23**, 1514–1521 (2013).
67. Weir, B. S. & Hill, W. G. Estimating F-statistics. *Annu. Rev. Genet.* **36**, 721–750 (2002).
68. Rosenbloom, K. R. *et al.* ENCODE data in the UCSC genome browser: year 5 update. *Nucleic Acids Res.* **41**, D56–D63 (2013).
69. Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222–226 (2012).
70. Tamura, K. *et al.* MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28**, 2731–2739 (2011).

Acknowledgements

We thank M. Phifer-Rixey and A. Werner for computational assistance during this project. W. Allen, K. Ferris and T. Hendry provided useful comments on earlier drafts. M.J.S. was supported by a Ruth Kirschstein National Research Service Award from NIH. M.W.N. was supported by NIH R01 GM074245.

Author contributions

M.J.S. conceived the project; M.J.S. and M.W.N. designed the study; M.J.S. collected and analyzed the data; and M.J.S. and M.W.N. wrote the paper.

Additional information

Supplementary Information accompanies this paper at <http://www.nature.com/naturecommunications>

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at <http://npg.nature.com/reprintsandpermissions/>

How to cite this article: Sheehan, M. J. and Nachman, M. W. Morphological and population genomic evidence that human faces have evolved to signal individual identity. *Nat. Commun.* 5:4800 doi: 10.1038/ncomms5800 (2014).